Adaptive Feature Selection and Risk Prediction for Mental Health Decision Support

Nawal Aly mahmoud Aly Zaher

Doctor of Philosophy

June 2018

Aston University

Adaptive Feature Selection and Risk Prediction for Mental Health Decision Support

Nawal Aly mahmoud Aly Zaher
Doctor of Philosophy
June 2018

Thesis Summary

Mathematical modelling of mental health risks is a laborious task, since assessment records contain diverse combinations of variables and a huge amount of missing data. In addition, risk judgements made by assessors are not clearly formulated from the available variables. The problem consists of two parts: first, selecting the most appropriate variables and, second, predicting risk using these variables, in real time and in a manner that is clinically explainable.

In this thesis, an adaptive feature selection algorithm is proposed based on Minimum Redundancy Maximum Relevance (MRMR) that builds on and extends the Dynamic Feature Selection and Prediction (DFSP) algorithm [1]. A feed forward approach is utilised to reduce computational complexity. The selected features are used to predict risk through a linear regression model. The predictions are linearly adjusted and unequal variance decision boundaries are used to handle heteroscedasticity. Two preprocessing steps are applied to reduce dimensionality and redundancy. The proposed algorithm is called Adaptive Feature Selection and Prediction (AFSP) and a method to autonomously update all its parameters, is devised.

When the algorithm is applied to suicide risk prediction, the results show that AFSP has better prediction accuracy than its predecessor, DFSP. The results also highlight the improvement in accuracy and/or speed introduced by each component of AFSP.

The algorithm is also applied to two sub-concepts within suicide risk. First, AFSP is used to determine whether the absence of current intention stated by patients is reliable or not. Second, AFSP is used to predict patients that are in an episode of clinical depression. The results of current intention and depression prediction are statistically significant.

The algorithm is intended to provide mental-health practitioners with prediction advice in real time, selecting the best factors for explaining predictions and updating parameters autonomously off line so that they reflect the latest data.

Keywords: Risk assessment, Suicide risk, Suicidal intent, Missing data.

Acknowledgement

First, I would like to thank my supervisor, Dr. Christopher Buckingham, whose guidance was imperative to the progression of this thesis. I would like to acknowledge everyone working on GRiST, for providing the platform and the data that facilitated this research. I am also grateful to Dr. Sherine Saleh for the work she did on risk prediction, which led the way to this research.

I would like to dedicate this work to my family. Mohammad, my husband, who has always been there for me and Adam, my son, who understood when mummy had to work. There are not enough words to thank my parents, who supported and guided me throughout my life. I would also like to thank my parents-in-law for helping out with Adam, when I had to work on my thesis. I am thankful for having two amazing sisters and a brother who always brought happiness into my life.

Finally, I would like to show my deep gratitude to my employer, the Arab Academy for Science and Technology (AAST), for sponsoring my PhD studies and giving me this wonderful opportunity.

Contents

List of Abbreviations

GRiST	Galatean Risk and Safety Tool
AFSP	Adaptive Feature Selection and Prediction
DSS	Decision Support System
ALERT	ALgorithms for Effective Reporting and Treatment
FACE	Functional Analysis of Care Environments
CARDS	Clinical Assessment of Risk Decision Support
GIRAFFE	Generic Integrated Risk Assessment For Forensic Environments
MG	Membership Grade
DK	Don't Know
RI	Relative Influence
DFSP	Dynamic Feature Selection and Prediction
LLBFS	Local-Learning-Based Feature Selection
MRMR	Minimum Redendancy Maximum Relevance
MRMRQ	Minimum Redundancy Maximum Relevance Quotient
VDM	Value Difference Metric
SA	Simulated Annelaing
GA	Genetic Algorithm
FFS	Floating Forward Selection
FBS	Floating Backeard Selection
MSE	Mean Square Error
SVM	Support Vector Machine
NN	Neural Network
DT	Decision Tree
HMM	Hidden Markov Model

ICA	Independent Component Analysis
DV	Dependent Variable
IV	Independent Variable
OLS	Ordinary Least Squares
MLR	Multinomial Logistic Regression
IRLS	Iteratively Reweighted Least Squares
GLM	Generalised Linear Model
GMM	Gaussian Mixture Model
ML	Maximum Likelihood
PCA	Principal Component Analysis
UVSD	Unequal Variance Signal Detection
MLE	Maximum Likelihood Estimation
MRMRD	Minimum Redundancy Maximum Relevance Difference
CFS	Correlation-based Feature Selection
RF	Random Forest
LR	Linear Regression
MRMRN	Minimum Redundancy Maximum Relevance Normalized
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
FPR	False Positive Rate
MDD	Major Depressive Disorder
DHT	Digital Health Technologies
RCT	Randomized Controlled Trial
QoS	Quality of Service
NIHR	National Institute of Health Research
GDPR	General Data Protection Regulation

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Mental Health Risks

Mental health risk refers to severe outcomes that result from mental health problems. Among those risks (outcomes) are self-harm, harm-to-others (including violence), self-neglect, vulnerability and suicide [11, 12]. The factors contributing to the development of mental health risks vary widely. Risk factors include, but are not limited to, history of mental illness (mental conditions and disorders), traumatic events, chronic and fatal health problems, history of self-harm or violence, history of suicide attempts, poor living conditions or mental health services, personal or financial problems, addiction, general behaviour and personal traits, feelings and emotions and history of abuse.

In recent years, the high frequency of suicide attempts and completed suicides [13] has drawn attention to risk assessment within the mental health domain. Mental health risk assessment is the process of determining a measure of both the probability and severity of one of the previously mentioned outcomes. Although mental health risks are diverse and equally alarming, suicide is the most rigorously researched risk [14].

Suicide risk assessment does not mean computing the probability of death by suicide, but is a measure of the intent and familiarity [15, 16] of a person with regards

to parasuicide (attempting suicide) [17] and completed suicide. On the other hand, while self-harm is treated as a separate risk, the degree of self-harm may indicate an actual risk of death, and thus self-harm is one of the factors that contribute to suicide risk [18].

1.1.1 Clinical Risk Assessment

Mental health risk assessment is most commonly carried out by clinicians. Clinical risk assessment involves quantifying risk based on a decision made by a professional. Clinicians make their evaluation of risk through asking a patient questions about probable causes and symptoms of a particular risk [19]. The choice and order of questions depends on the clinician's intuition and expert opinion. The final risk judgement is based on the clinician conception of the answers provided by a patient [20]. If a patient presents high risk, a management plan is usually devised by the clinician [21].

1.1.2 Statistical Modelling of Risk

Statistical models may be used to map answers in an assessment to risk categories. Through training, relationships between the data in an assessment and the outcome, may be deduced. These models are usually developed from clinical expertise [22] and are not meant to replace clinical assessments but rather complement them. The risk model may be used to guide para-professionals through questions, to minimise the set of information required to compute the risk prediction [23].

1.1.3 GRiST

The Galatean Risk and Safety Tool (GRiST) [24] is a web-based platform that is used for risk assessment by clinical experts and members of the public in the UK. It provides an adaptive interface for mental health risk assessment and management

[25]. The population in GRiST includes several age groups: children, adolescents, adults and older adults.

GRiST represents mental health risks through a mind map. The map or "tree" for each risk consists of higher-level concepts such as history of suicide, current intention and, feelings and emotions. These are divided into lower-level concepts and/or leaf nodes. The structure and scope of GRiST will be discussed in detail in Chapter 2, since it is the source of all data used in this research.

1.2 Limitations

Several problems face mental health risk assessment, the most common of which are listed below.

1. There are large variations within the population for a single risk. For example, patients with no declared suicidal intent are usually perceived as low risk, which affects the motivation for asking additional questions in other areas such as their general feelings and emotions. These differences require a tailored treatment for each individual case.

2. Parts of the data may be missing for various reasons, discussed in Chapter 2. Missing data is a limiting factor for statistical modelling of risk.

3. Records may contain redundant information, since questions asked during an assessment are usually related through a hierarchical structure.

4. Risk judgements made by clinicians are subjective, as these are based on intuition and experience.

1.3 Thesis Objectives

This thesis describes the development and implementation of the Adaptive Feature Selection and Prediction (AFSP) algorithm, to handle prediction from incomplete high-dimensional subjective data sets. Although AFSP is applied to mental health data from GRiST, the algorithm is applicable to other data sets and/or domains. The objectives of the thesis are as follows;

- Select the most relevant set of features to explain the risk.
- Reduce redundancy within the selected feature set.
- Provide accurate predictions of clinical risk judgements.
- Reduce dimensionality and complexity to facilitate real-time implementation.
- Produce a machine learning algorithm that can be autonomously updated as new data arrives and that can provide real-time predictions during clinical risk assessments as part of an intelligent Decision Support System (DSS).
- Apply the algorithm to diverse problems to ensure its general applicability.

1.4 Organization of Thesis

This thesis is composed of 8 chapters, detailed in the list below;

- Chapter 2 introduces the GRiST DSS and reviews feature selection and risk prediction techniques that may handle data-related issues within the domain.
- Chapter 3 presents the rationale and details of AFSP and the modifications implemented to overcome data-related issues.

- Chapter 4 discusses the application of AFSP to suicide risk prediction and details the parameters needed for implementation.

- Chapter 5 explains the results of suicide risk prediction using AFSP in comparison to other methods.

- Chapter 6 extends the applicability of AFSP to predict influential sub-concepts within suicide risk, namely, current intention of suicide and clinical depression; and highlights the superiority of using AFSP compared to a fixed feature set.

- Chapter 7 discusses how AFSP may be deployed within GRiST in clinical practice and the monitoring and maintenance needed to operate it.

- Chapter 8 summarises the research undertaken and proposes possible future modifications of AFSP to overcome its current limitations.

Chapter 2

Background

2.1 Mental Health Risk Assessment

Risk assessment is the judgement people make about both the probability and severity of an event [26]. Mental health risk assessment addresses risks originating from mental health problems. The accurate identification of risk and risk factors is imperative to clinical decision making and risk management. The aim of mental healthcare is to minimise risks for the well being of the people at risk and their community. An efficacious mental healthcare system has to be able to identify risk factors, assess mental health risks, devise and implement management plans and assess outcomes [27]. Due to the intricate nature of mental health risks, healthcare systems have to iteratively perform the previously mentioned tasks, to ensure the effectiveness of risk management plans through re-evaluating risks and adjusting the plans to minimise them [2]. Therefore, not only does risk assessment pave the way for risk management but it is also continuously required to improve management plans [28], as shown by Figure 2.1.

Figure 2.1: Risk management cycle, reproduced from [2]

2.1.1 Clinical Assessment

Clinical risk assessment is the process through which clinicians evaluate risk from risk factors obtained by interviewing patients. Many attempts have been made to unify and formalise the factors contributing to various mental health risks [29, 30, 18, 31], yet the influence of these factors on risk remains highly subjective to the assessor. A clinician usually selects cues that are most relevant to the risk and determines their relative influence on risk, based on experience and intuition [32]. Hence, risk judgements from several clinicians may be different for the same patient. In many cases, different clinicians may acquire different parts of the data, depending on which factors they believe are important, as questions asked during clinical assessments are usually directed by the patient's responses.

2.1.2 Data Collection Tools

An aim to regularize clinical risk assessments is by using data collection tools. These tools intend to provide a structure and highlight the risk factors relevant to certain risks. The structure and factors are usually inspired by expert knowledge of the domain, as these tools are designed to meet the clinicians' need for an organised platform for collecting data. However, data collection tools do not offer any decision support advantage.

2.1.3 Decision Support Systems

Decision Support Systems (DSSs) use risk assessment tools to compute mental health risks through mathematical or logical models. DSSs are usually based on data-driven models or knowledge-based models. Both categories utilise mathematical models to predict mental health risk from cues pertaining to the risk, but through different approaches. Data-driven models try to find patterns in the data, whereas expert-based models depend on a structure laid down by experts in the risk domain.

The current availability of tools to assess various mental health risks is limited [33]. The main reason for lacking a comprehensive tool for mental health risk assessment is that tools have been designed for only specific users and/or risks. One of the very first tools designed for self assessment is the Self-Rating Questionnaire [34], that aims to replace interviews with clinicians and provide a fast inexpensive method of risk assessment. Other tools were designed to facilitate data collection and analysis by para-professionals [35] due to their lack of experience in the domain. DSSs such as the ALgorithms for Effective Reporting and Treatment (ALERT) [36] were designed to cross-validate self-reported risks with clinical judgements, yet little evidence exists on the validity of ALERT. Functional Analysis of Care Environments (FACE) [37], on the other hand, has good validity and targets several risk areas [2], but users are required to pay for a license and get training to use the software. Diversely, some tools were designed to target only specific risks within the mental

health domain, such as the Clinical Assessment of Risk Decision Support (CARDS) [38] which is used for the assessment and management of risk of violence only. Other tools were developed for a specific type of patients, such as the Generic Integrated Risk Assessment For Forensic Environments (GIRAFFE) which addresses mental health risks within forensic psychiatry only [39].

The absence of a DSS covering a variety of mental health risks that can accommodate the needs of the clinicians and the domain is the main reason behind the limited deployment of DSSs in mental health [40]. GRiST [24] addresses these issues by building an ontology and interface based on clinical expertise.

2.2 GRiST

GRiST [24, 41] is a web-based DSS for mental health risk assessment, that is based on knowledge structures used by mental health practitioners [32, 42, 43], obtained by interviewing qualified clinicians [22, 44]. Expert knowledge collected through interviews was combined to generate a tree-like structure using the risk factors informed by the clinicians [45]. Feedback from practitioners using GRiST has been used to continuously modify the system to meet the service user needs and expectations. GRiST service users include NHS services, private hospitals, charities, and members of the public [46]. The database has more than 250,000 completed assessments for over 100,000 patients provided by 3,000 mental-health professionals. On-going research ensures that GRiST is being continuously improved. The following subsections detail several aspects of GRiST.

2.2.1 Objectives

The objectives of GRiST are concisely listed below:

1. Provide an interactive platform for clinicians to conduct assessments

2. Make risk predictions that conform with clinical risk judgements

3. Explain risk factors for a better management of risks

4. Provide an easy-to-use platform for self-assessment [46]

5. Improve mental health quality

2.2.2 Ontology

The user interface for GRiST, shown in Figure 2.2, takes the form of a mind map. The structure clearly divides patient data into risk-specific, and general questions (e.g. questions about personality, state of mind, behaviour, etc...) that may pertain to any risk [47]. The mind map makes it easy for clinicians to choose which area of concern they would like to address in a compact and straightforward manner. The structure makes it possible for clinicians to fill only the details they think are necessary or relevant to a patient and allows clinicians to choose the order in which they conduct an assessment, instead of using a sequential must-go-through-it-all interface [25, 41].

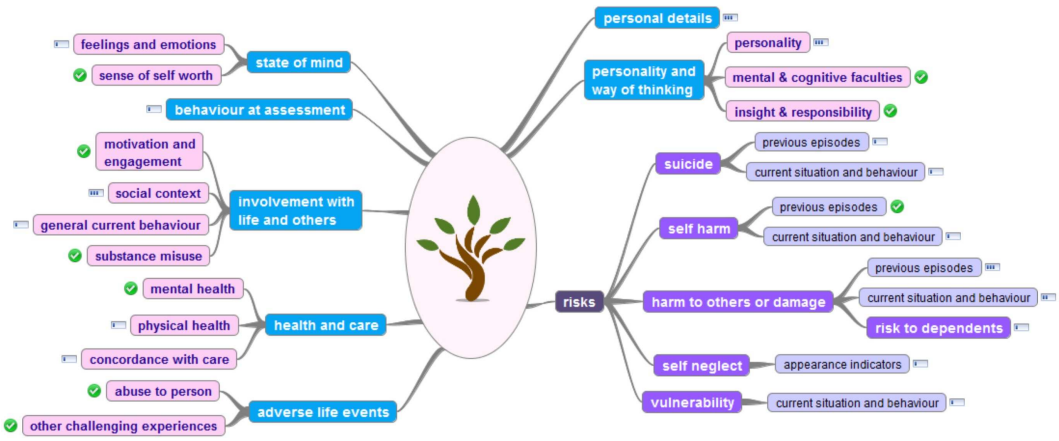

Figure 2.2: Snapshot of the user interface

Each node of the mind map is further divided into more detailed questions. Two types of nodes (questions) exist: concept nodes which constitute filter questions and leaf nodes which are non-filter questions. Filter questions open up a number of

subsequent filter or non-filter questions that provide more detail about the concept represented by the filter. Figure 2.3 shows an example of a concept question and Figure 2.4 shows the subsequent questions when the filter is answered “Yes”. If an area is of no concern to the assessor or is irrelevant to the subject of the assessment, the corresponding concept question is answered as “No”, and thus the underlying questions automatically become irrelevant to the assessment and are not asked.

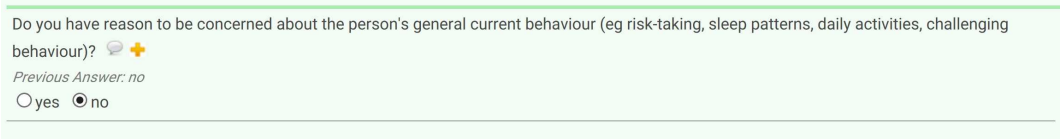

Figure 2.3: A concept question answered “No”

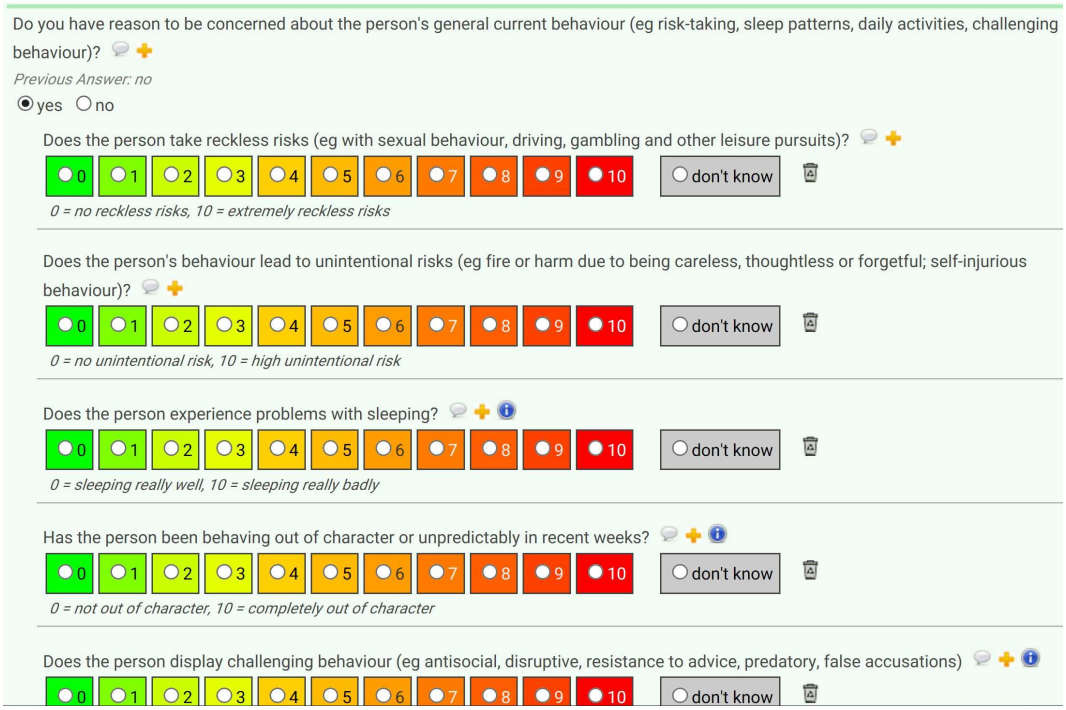

Figure 2.4: A concept question answered “Yes” and subsequent nodes

2.2.3 Numerical Representation

A patient’s answer to a question during an assessment is interpreted within GRiST as a Membership Grade (MG) ranging from 0 to 1. Concept questions (sometimes referred to as parents) have three possible answers, “Yes”, “No” or “DK” (“Don’t Know”). A “No” answer is given an MG of 0 and this denotes the least influence to

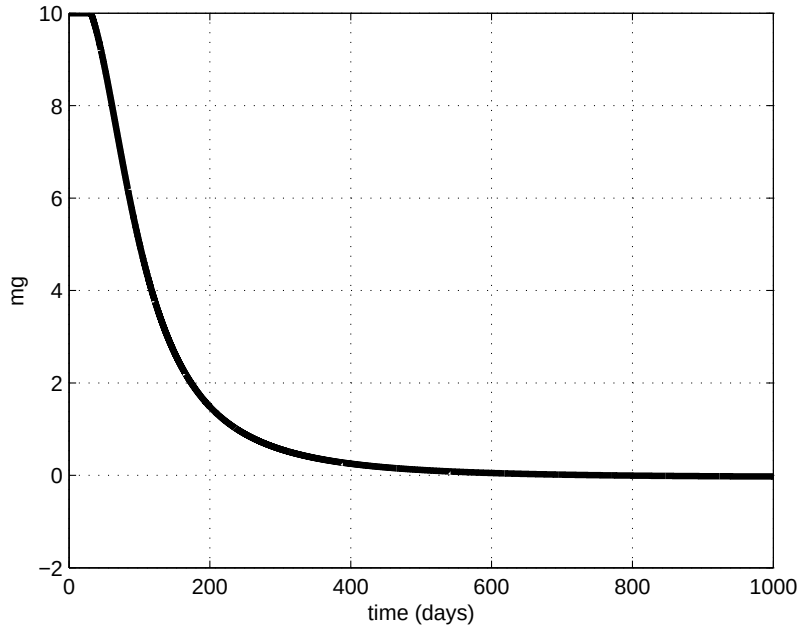

Figure 2.5: The MG value of the most recent episode cue against the time lapse since the most recent attempt

a risk, while an answer of “Yes” gives an MG of 1 and is the highest influence a cue may have on the risk. Non-filter questions (leaf nodes or children) have a graded answer on a scale of “0” to “10”, which is directly scaled into MGs from 0 to 1 and are interpreted as lowest to highest influence of a cue on the risk, but may also be answered as “DK”. Chronological cues, that are usually reflected as dates in GRiST, are transformed into MGs using a mapping function. Figure 2.5 shows the mapping between the time lapse since the latest suicide attempt by a patient and the MG value of the most recent episode cue.

An answer of “DK” or no answer at all results in a cue being marked as unavailable, to distinguish between: an answer of “No” or “0”, and a missing cue. A value of 0 holds information about lack of risk from a certain cue, while an unanswered cue should not contribute to knowledge about risk and should not be assumed as 0. This distinction between missing and “No” answers is important as it reduces the assumptions one may make about the data.

Clinical risk judgements for each assessment are provided by clinicians. The risk judgements are given on a scale of “0” to “10”, with “0” being no risk and “10”

being highest risk. These are also scaled to values 0 to 1 within GRiST and provide targets (labels) for the modelling and prediction of risk within GRiST.

2.2.4 Risk Factors

There are several factors that may affect mental health risks. Among the questions in GRiST there are risk factors, that when answered in the affirmative would result in higher risk of mental health risks. On the other hand, there are protective factors or desirable traits, whose existence would attenuate the risk and help in preventing undesirable outcomes.

Examples of risk factors are more obvious in the case of mental health risks, such as attempted suicide or self harm, history of abuse or depression. However, protective factors such as a good social context or being married or with partner are also addressed during assessments and incorporated in the final risk judgement. These factors influence clinical risk judgements and are coded within GRiST such that they reduce the risk prediction such that, a low mg value is attributed to the cue when it has a positive effect (reduces risk).

The cues are also labelled with padlocks that indicate whether or not the cues may change value and if they do, how likely are they to change. The symptoms of risk episodes, such as suicidal ideation or intent are more likely to change values from one assessment to the next (No padlock), compared to contextual variables, such as the nature of suicide attempts, that can change values, but not very often (Silver padlock). On the other hand, historic factors such as whether or not the person has ever attempted suicide may change only once from “No” to “Yes” but may never change again (Gold padlocks).

2.2.5 Diagnostic Validity

Although GRiST does not generate automated predictions, there are several advantages of using GRiST instead of simple clinical interviews. First, the construct of the risk with leaf nodes and filters and their order and placement in the tree is a result of clinical consensus [45], which reduces bias introduced by the order and choice of questions within the assessment. Second, GRiST verifies patient data against previously collected data for the same patients to avoid illogical changes for monotonically changing variables, such as the number of suicide attempts. Moreover, previously collected patient data is available for the assessor to check at any time in an accessible manner, and any static data that has been previously collected such as gender, age or ethnicity is automatically filled in by GRiST which reduces the amount of data the clinician has to gather and enter during each assessment.

Since the structure and order is modified through feedback from users and updates have been implemented following service users' feedback [46], the friendliness of the platform to practitioners is always improving. The ease of use has motivated the development of a self-assessment tool, myGrace [24, 46], based on GRiST structure. The list below summarises some of the advantages of GRiST [24]:

- links low-level cues, through higher level concepts, to mental health risks
- provides a formal structure and location for each piece of patient data
- acts as an index to data held in other patient documentation (assessments of other risks or previous assessments) to facilitate its verification and linkage
- has potential to populate information in patient's previous records to avoid double data entry
- makes it easy to find and format information in structured and organised reports

2.3 Issues with Analysing the Data

A mental health risk prediction algorithm encounters many data-related issues, including high dimensionality, irrelevance, redundancy, subjectivity and missing data. The reasons behind each of the previously mentioned issues and their possible effects on risk prediction is discussed in the following subsections with examples to highlight the differences between them.

2.3.1 High Dimensionality

The causes and symptoms of mental health risk vary widely [11, 18, 30, 48, 49], which broadens the scope of mental health data. Moreover, other pieces of information corresponding to personality, feelings and emotions, living conditions, general health and state of mind, are indirectly involved in risk assessment. As an example, for suicide risk assessment in GRiST, 177 different cues are included in risk assessment, of which 30 cues are directly related to suicide and 147 cues relate to more general mental-health and well-being of the patient. As the number of features increases, the volume of the space increases and the number of training samples, required to find a model that generalises adequately, increases exponentially [50]. Even with 250,000 training records, the data will cover a small fraction of a 177-dimensional space, this is often referred to as the curse of dimensionality [51]. The curse of dimensionality would also affect the choice of a prediction algorithm, as different model assumptions require different amounts of data during training and prediction [52]. This will be discussed in Section 2.5.4.1 and the amount of data required for training and prediction will be compared for different prediction approaches.

2.3.2 Irrelevance

The relevance of a cue to a specific mental health risk is established by experts in the domain, in the general sense. However, the relevance of a cue to the risk for a specific patient is defined by the patient's answer and the feature's applicability to the patient's case. For instance, if a patient answers "No" to having ever attempted suicide, then questions about their previous suicide attempts become automatically inapplicable. Another example is when patients have the same answer for one cue, but because of the absence/presence of other pieces of information (that may differ among patients), that one cue could have relatively different importance among different patients, and thus may be highly relevant to the risk in one case but not in another.

2.3.3 Redundancy

The nature of the domain incites redundancies among different cues. One way redundancies could arise is when a certain answer to a question always means that another has been answered in a specific manner. An example of this is a filter question, where any children holding any value would necessitate that the filter question value is 1; and thus the conditional probability of a value of 1 for a filter question, given a child has been answered, is 1, regardless of the child's MG value. Other more subtle redundancies may exist that are dictated by the nature of the data. A large redundant feature set induces unwanted complexity that is not informative and may also affect predictive performance negatively, if the prediction model assumes that features are independent [53].

2.3.4 Subjectivity

The way clinicians conduct assessments and evaluate risk is highly subjective. When conducting an assessment, a clinician would choose to ask questions that are most relevant to the risk being assessed from the available set of cues [32], and would

shift focus from one variable to the other depending on the patient's answer and on intuition [19, 54]. The way clinicians do this is, unfortunately, ambiguous and is not formally defined for a DSS to follow. Consequently, the choice of which questions to be asked and which cues are present in an assessment record depends on clinicians and what they think is relevant in a particular situation. Recent studies suggest that clinicians' gender and/or level of expertise and training will affect the way they perceive risk and conduct assessments [55, 56].

2.3.5 Missing Data

Certain parts of the data may not be collected if: the clinicians do not ask certain questions, patients are not willing to answer a specific question or when the data does not apply. Although missing data may be a result of one, all or none of the above issues (sections 2.3.2, 2.3.3, and 2.3.4), it poses a problem to data analysis and risk assessment, that is, unfortunately, too big to be overlooked [57]. In suicide risk assessment data, from GRiST, the average record has only 67 features available out of 177 possible features, which means that around 62% of the data is absent. One possible solution is to impute the missing data, but performing risk assessment using large amounts of fabricated mental health data (actually more fabricated data than actual data, in this case) is highly undesirable [58]. Whether the answer is "DK" or no answer is given at all makes a difference to the reason why the data is missing. However, the fact remains that these parts of the data, for which the value is unknown, cannot be used for calculating or explaining the risk, to maintain a high level of confidence in risk prediction.

2.4 Feature Selection Techniques

The aforementioned data-related issues call for using a feature selection technique that does not only find complete sets, but also improves the quality of the set in terms of reducing redundancy and maximising relevance to the risk. The advantage

of using a feature selection technique is that mental health data contains many features that are redundant and/or irrelevant, and thus their removal would, at the very least, reduce complexity without incurring any losses in information.

There are three possible approaches to feature selection: filter, wrapper and embedded methods [59]. Filter methods use a scoring scheme to rank variables based on a metric such as distance, information gain, dependency or similarity [60]. The scores are a reflection of the relationship between the candidate features and the variable to predict, regardless of the prediction model, which makes filter methods computationally inexpensive and immune to over fitting. Wrapper methods, on the other hand, are model-specific as these perform predictions at each round to score candidate feature sets. Each new set is used to train a model and the error rate of the model gives the score for that candidate feature set [61]. Wrapper methods are computationally intensive for a large number of candidate features and are more susceptible to over fitting, but usually provide the best performing feature set for a particular type of model. Embedded methods are logically the same as wrapper methods, but computationally optimised by performing feature selection and classification simultaneously to take advantage of any computations performed during selection that may be used for classification [62]. A comparison by [63] recommended the use of filter over wrapper methods for their simplicity and generalizability. However, filter methods do not, traditionally, deal with redundancy issues. Redundant features do not introduce any additional information and may need to be removed to speed up the learning algorithm and boost performance [64].

2.4.1 Selection Criteria

Examples of the metrics used to score features in filter methods are correlation which is a statistical measure of dependency, mutual information which is an information based metric and distance based metrics such as Relief and Local-Learning-Based Feature Selection (LLBFS).

2.4.1.1 Correlation

Correlation is any of a broad class of statistical relationships involving dependence, though the most commonly used measure of correlation is the Pearson correlation coefficient, which measures the linear relationship between two variables. The correlation coefficient ρ_l between the target of a prediction Y and a possible predictor X_l is given by 2.1. If two variables are independent, their correlation will be zero but the opposite is not true as correlation only measures linear dependency.

$$\rho_l = \frac{cov(X_l, Y)}{\sqrt{\sigma_l^2 \sigma_Y^2}} \tag{2.1}$$

where $cov(X_l, Y)$ is the covariance of X_l and Y, and σ_l^2 and σ_Y^2 are the variance of X_l and Y, respectively.

2.4.1.2 Mutual Information

Mutual information I_l measures the amount of dependence between two random variables X_l and Y using 2.2 [65].

$$I_l = \sum_{y \in Y} \sum_{x_l \in X_l} p(x_l, y) log\Big(\frac{p(x_l, y)}{p(x_l)p(y)}\Big) \tag{2.2}$$

Contrary to Pearson's correlation, mutual information does not measure linear dependency between variables, but is an entropy-based information metric. The entropy $H(Y)$, in Equation 2.3 is a measure of the minimum number of digits that can represent all possible outcomes of a random variable Y. The conditional entropy $H(Y|X_l)$, in Equation 2.4, measures the number of digits needed to represent Y given that the value of X_l is known, and thus represents the amount of uncertainty remaining in Y when X_l is known. Mutual information I_l may be calculated by Equation 2.5 as the difference between the total uncertainty of Y, which is represented by $H(Y)$, and the remaining uncertainty in Y when X_l is determined, which

is represented by $H(Y \mid X)$. Hence, I_l measures the information gained about Y given the value of X_l [66].

$$H(Y) = -\sum_{y \in Y} p(y) log(p(y)) \tag{2.3}$$

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} log(p(x, y) p(y \mid x) \tag{2.4}$$

$$I(Y, X_l) = H(Y) - H(Y \mid X_l) \tag{2.5}$$

2.4.1.3 Relief Based Feature Selection

Relief-based feature selection encompasses a class of algorithms based on Kira and Rendell's [67] Relief algorithm, that is used to evaluate features in binary classification problems based on the distance between a sample and the nearest hit and miss. ReliefF [68] is an extension of Relief to multi-class problems, by searching for k nearest hits and misses from each different class and averaging their contributions for updating the scores, weighted with the prior probability of each class. Whereas, Regression ReliefF (RReliefF) [69] is an update of ReliefF making it suitable for regression.

The basic idea of Relief-based feature selection is to score predictors based on how well they distinguish between close instances of different classes [70]. Since mental health risk assessment is a multi-class problem, we shall focus on ReliefF and RReliefF. The feature score S_X for ReliefF is given by Equation 2.6 [70]. The approximation of the probabilities are based on the difference function $diff(X, C_i, C_j)$, represented by Equation 2.7 [60, 70], which denotes the distance between instances i and j with regard to dimension(feature) X.

$$S_X = p(X_i \neq X_j \mid i \neq j) - p(X_i \neq X_j \mid i = j) \quad \forall i, j \tag{2.6}$$

where $p(X_i \neq X_j \mid i \neq j)$ and $p(X_i \neq X_j \mid i = j)$ are the probabilities of a different value of X given that instances I_i and I_j belong to different classes or the same class, respectively.

$$diff(X, I_i, I_j) = \frac{|X_i - X_j|}{X_{max} - X_{min}} \tag{2.7}$$

In ReliefF each of the k nearest hits and misses equally contributes to the probabilities in Equation 2.6, which is equivalent to a weight of $1/k$ assigned to each instance [69]. Whereas in RReliefF the hits and misses are assigned different weights $d_k(i, j)$ (Equation 2.8) depending on their rank $rank(I_i, I_j)$.

$$d_k(i, j) = e^{-\frac{(rank(I_i, I_j))^2}{\sigma^2}}, \tag{2.8}$$

where σ is a user-defined parameter.

2.4.1.4 Local-Learning-Based Feature Selection

LLBFS [71] utilizes local learning to divide a globally non-linear problem, with no assumptions about the underlying model, into locally linear ones. Calculating feature scores is then transformed into an optimization problem formulated as a logistic regression problem, that is penalised to encourage sparseness. This method works well when there is a large number of irrelevant features [71].

2.4.2 Minimum Redundancy Maximum Relevance

Unlike the previously mentioned selection techniques, Minimum Redundancy Maximum Relevance (MRMR), introduced in [72], tries to balance the relevance to the predicted variable with redundancy within the feature set. MRMR is based on maximising a relevance term D, which represents the amount of information between a predicted variable and a feature, and minimising a redundancy term R, which represents the amount of information shared by variables within a set. Originally, a

difference between D and R was used to combine both criteria and the feature set that maximised the difference ϕ in 2.9 was chosen.

$$\phi = D - R \tag{2.9}$$

However, the redundancy and relevance terms are not comparable in many cases, which may result in one overpowering the other in the selection process [73, 74]. A scaling parameter is suggested by [73] to balance both terms, but the parameter is manually chosen and the authors do not give any evidence of how parameters are chosen. In [74], mutual information is averaged by the number of features in the set. This algorithm assumes knowledge of the optimum number of features beforehand, which is not usually the case. Another implementation of MRMR, is through maximising a quotient score Q in 2.10 instead of difference to overcome normalization issues [75]. This is usually referred to in the literature as MRMR Quotient (MRMRQ) [76].

$$Q = \frac{D}{R} \tag{2.10}$$

Several parameters may be used for measuring relevance and redundancy. When MRMR was first introduced, [72] suggested using mutual information for calculating both terms. A survey by [75] compared several redundancy and relevance measures within the context of MRMR, including rank correlation, conditional distributions and Value Difference Metrics (VDMs).

In mental health risk assessment, the need to select from a huge set of candidate features may contradict the feasibility of implementing MRMR, as enumerating all possible combinations may be computationally prohibitive. For finding L features from M possible candidates, one needs to repeat the computations $\binom{M}{L}$ times [3], which results in factorial class of computational complexity, given the length of the feature set L is known. Often, one has no prior knowledge of the number of variables that would best explain the predicted variable, and thus would need to repeat the selection process with different numbers of features which further

increases the computational complexity. A feed forward approach, illustrated in Figure 2.6, was suggested by [3] to speed up the search process by maximising the objective function incrementally only and devising a stopping condition instead of using a fixed number of features. The selection stopped when the change in Mean Square Error (ΔMSE), computed by performing predictions on the data, was lower than a predetermined error threshold ϵ. However, this feed forward approach turned the selection process into a wrapper method.

Figure 2.6: Feed forward approach to MRMR, reproduced from [3]

2.4.3 Search Strategies

Although the forward selection approach [3] seems appealing, it is not the only plausible replacement for the brute-force implementation of MRMR. Search strategies may be divided into three categories: exponential search, sequential search and randomized search. Exponential search strategies, such as Branch and Bound [77], are computationally expensive, since their complexity increases exponentially

with the number of features. Examples of sequential search strategies are forward selection, backward elimination [78], bidirectional search [79] and floating search [79], while Simulated Annealing (SA) [78] and Genetic Algorithms (GA) [80] are examples of random search strategies. Randomized search usually outperforms sequential search methods in terms of the quality of the features but not in terms of speed [78]. Hence, we will consider only sequential search techniques for the sake of reduced computational complexity.

2.4.3.1 Forward Selection vs Backward Elimination

In forward selection, one starts with an empty set and adds the temporally optimum feature at each iteration. Once a feature is added, it may never be removed, and thus, each feature affects future choices but does not back propagate its effects. On the other hand, backward elimination starts with the full set of features and removes the worst feature (one with the minimum score) at each iteration. Once a feature is removed it may never be allowed back into the feature set [78]. In both cases the algorithm stops when a predetermined condition is met, such as a fixed number of features, accumulated score threshold or an error threshold. Evidently, forward selection is generally faster than backward elimination when the feature set size is small compared to the candidate set and vice versa.

2.4.3.2 Bidirectional Search

Bidirectional search capitalizes on the advantages of both forward selection and backward elimination, by taking turns performing forward and backward rounds [79]. To avoid infinite loops, features added in the forward round are never removed in later backward rounds and vice versa. For sets that are not too small or too large, bidirectional search will reach a solution faster than both forward and backward searches, since the number of candidates is reduced at each round of forward selection by previous rounds of backward elimination (since eliminated features may not be added), and the number of features to be considered for elim-

ination in backward rounds is reduced by the number of forward selections previously performed (since those features may not be removed). The complexity of a bidirectional search is $O(2M^{L/2})$ compared to a complexity of $O(M^L)$ for forward selection or $O(M^{M_{max}-L})$ for backward selection, where M is the average number of candidates at each round (since it is not constant), L is the length of the selected feature set and M_{max} is the total number of features to be considered [81].

2.4.3.3 Floating Search

There are two possible setups for floating selection, Floating Forward Selection (FFS) and Floating Backward Selection (FBS). They differ in the order and preference of one step over the other. In FFS, after each forward step, backward steps are performed as long as the objective function increases. On the other hand, FBS performs forward steps after each backward step as long as the objective function increases [79]. Note that floating selection does not work for monotonically increasing objective functions and that a degree of book keeping has to be maintained to avoid infinite loops. While floating selection does seem more demanding than other sequential algorithms, it has the benefit of some back tracking capabilities [82].

2.4.4 Fixed vs Dynamic Sets

Feature selection techniques are used to select the best features, within the context of a problem, in order to reduce dimensionality and/or enhance predictive performance. However, when feature selection is applied on the entire population at once, the resulting feature set will be the same for all cases (i.e. fixed). The problem with a fixed set is that it will not work well when many instances of the features are missing and with different patterns of missing features across the population. One way of dealing with missing data is case-wise deletion, which omits all cases with incomplete feature sets [83]. For mental health data, where a large percentage of the data is missing, case-wise deletion would render most of the assessments inadmissible for prediction, since they will not contain the complete feature set.

On the other hand, feature selection may be applied in a case-based manner [84], to select the best features, given the problem and the available set of features for each case individually. In this manner, feature selection starts off with the set of features available for a patient and selects a subset of those features to be used for risk prediction for that particular case. This method generates a dynamic set of features because it may be different among the different cases. The upside of a dynamic feature set is that it addresses the problems of missing data, irrelevance and subjectivity. The downside, however, is the need to perform feature selection for each and every assessment and to train a new model for prediction every time a new feature set is chosen. This may not be a problem if there are only a few subsets of the main population and a limited number of fixed feature set variants.

In metric based selection, where correlation, distance or information gain is used to score feature with relation to the outcome or the class, the order in which the features are selected will not change and thus selection would be a matter of sorting the available features based on the predetermined rank. However, when feature interactions are considered, such as in MRMR, the choice of feature will not only depend on its relationship to the response, but also on the degree of redundancy it introduces to the set, thus making the chosen set adaptive.

2.5 Prediction and Classification Techniques

In the mental health domain, prediction (as in dealing with future uncertainty) is the common notion attached to risk and is used interchangeably with classification. In statistics, prediction and classification may refer to different problems. While prediction does not necessitate discretisation of the output variable, classification must have a discrete outcome. In some sense, classification may be thought of as the prediction of the output to be one of K distinct classes $C_1, C_2, ..., C_K$. Hence, classification is a subset of prediction but the converse is not true.

Mental health risk clinical judgements are usually discrete in nature and may be handled as a prediction or a classification problem. Prediction would provide a

flexible outcome that is able to highlight differences between two cases even if they are members of the same class, as they are not necessarily identical and may have different membership grades to their class. Classification, on the other hand, offers more constrained outputs and may be harder to implement for high dimensional data, but a discrete output would resonate with clinical risk judgements.

Although it may be good practice to separate feature selection and prediction for high dimensional data, the need for feature selection is usually called for by the domain and not the prediction algorithm. Some prediction and classification methods may work with missing instances, while other algorithms have inherent feature selection capabilities. Hence, in the following subsections several prediction and classification techniques are reviewed, regardless of the need to perform feature selection beforehand.

Generative classifiers learn the joint probability $p(X, Y)$ of the inputs X and desired output Y and use Bayes rule to calculate the posterior probability $p(Y \mid X)$. Discriminative classifiers either calculate the posterior probability $p(Y \mid X)$ or a direct mapping function between the input and the output [85]. Generative techniques are often referred to as parametric techniques, while discriminative techniques are considered non-parametric, with regard to parameters of a statistical model. Generative classifiers make more assumptions about the underlying statistical structure than discriminative classifiers and are more computationally expensive to train. Support Vector Machine (SVM), Neural Network (NN), Decision trees (DTs) and regression analysis are discussed as the most common methods of discriminative techniques, while Hidden Markov Model (HMM) and Independent Component Analysis (ICA) are given as examples of generative models.

2.5.1 Support Vector Machines

SVM projects data onto a high dimensional feature space and searches, in that space, for the optimal hyperplane that separates two classes where the distance between members of different classes is maximum, using support vectors to describe the

hyperplane [86]. Generally speaking, SVM provides superior classification performance [87], compared to other approaches. However, the choice of kernel function used to project the data highly affects performance. Moreover, SVM is designed to solve binary classification problems only and is computationally expensive for high dimensional data [88].

2.5.2 Neural Networks

Despite NNs ability to handle noisy data during training and missing data during classification, NNs have several disadvantages. First, the number of hidden layer nodes is empirically determined and has an immense effect on performance accuracy and speed [89]. Second, its training or learning process is very slow, almost 40 to 100 times slower than regression (depending on the data) [90]. Third, NN lacks explanatory power, as the model involves hidden layers and, as a result, NN models are often considered black boxes. In mental health risk assessment explainability is highly important and using NNs will make it difficult to understand how decisions are made.

2.5.3 Decision Trees

The idea behind DTs is to iteratively partition the data based on feature values, until the subset of the population at each leaf is of the same class [91]. The top of the tree represents the root node (the entire population) and leaf nodes represent the classes. The major advantage of DTs is the visualization of data, which highlights the contribution of each node (feature) to the root (risk). However, as the dimensionality of the problem increases and variations within the population emerge, DTs may become too dense to be meaningfully visualised [88].

2.5.4 Regression Analysis

Regression involves directly mapping the output to the inputs using correlation between the output, usually referred to as the Dependent Variable (DV), and different input variables, referred to as Independent Variables (IVs) [92]. The DV and IVs can be either discrete or continuous or a mixture of both, depending on the nature of the data.

In multiple linear regression, the relationship between each IV and the DV is represented by a straight line (weight) and the DV is unbounded and continuous. Linear regression models are fast to compute and directly show the linear effect each variable has on the output prediction. While linear regression does not make an assumption about the underlying distribution of the variables, the optimality of Ordinary Least Squares (OLS), used for calculating the weights, can only be proved in the context of Gaussian random variables [93]. The linear model makes a huge assumption about the model structure and thus results in stable but possibly inaccurate predictions. The coefficients W are computed directly from the predictors X and the responses Y by Equation 2.11.

$$W = (X^T X)^{-1} X^T Y \tag{2.11}$$

The matrix inversion has a complexity of $O(L^3)$ where L is the number of IVs, and the matrix multiplication is $O(L^2 N)$, where N is the number of samples. Since N is often much larger than L, then the computational complexity of computing regression wieghts is in the order of $O(L^2 N)$.

Multinomial Logistic Regression (MLR) [94], in contrast to linear regression, can handle a categorical DV. MLR solves the classification problem using the logit function to predict the probability of belonging to a certain class. It is usually used with non-ordinal categories, where OLS methods cannot be used. MLR may be represented as solving $K - 1$ binary regression problems for K classes and thus it employs iterative methods [95] such as generalized iterative scaling [96], and Iteratively Reweighted Least Squares (IRLS) [97] as opposed to linear regression,

where coefficients can be directly computed from training examples. Since linear regression is a prediction algorithm, its complexity will not be affected by the number of classes. On the contrary, MLR has a computational complexity in the order of $O(KL^2N)$ [97], where K is the number of classes. In addition, if IRLS is used to train the model an additional $i_{max}L^2$ computations are needed to recalculate the weights at each iteration i, where i_{max} is the maximum number of iterations (user defined parameter). Moreover, MLR performance is inconsistent among classes, as it tends to underestimate probabilities for rarely occurring classes [98]. This is particularly problematic for high risk cases, that are critical but have a low occurrence frequency.

Generalised Linear Model (GLM) is the general case of linear regression, where the relationship between the DV and the IVs is not linear. GLM uses a link function, such as log or identity, to transform the relationships. The choice of the link function depends on the underlying assumption of the joint probability distributions of the IVs with the DV [99]. Similar to MLR, IRLS is used to train a GLM and its computational complexity is in the same order of MLR [100].

2.5.4.1 Local Methods vs. Least Squares

Another approach to prediction is the k-nearest neighbour approach, which computes the prediction Y as a linear combination of the nearest neighbours by uniformly averaging the outcome y_i of the k nearest neighbours x_i to a test vector x, as in Equation 2.12 [52].

$$Y = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2.12}$$

where $N_k(x)$ is the neighbourhood of x defined by the k closest points x_i.

Although local methods make very little assumptions about the data, they are highly unstable. Their instability is manifested in high dimensional data, since it gets harder to find k uniform neighbours in a larger space. Particularly, to cover a fraction r of the volume of the space, one would need to cover the range $R_L(r) = r^{\frac{1}{L}}$ for each input variable, where L is the number of input features [52]. This will be

harder to achieve if the data is sparse, which is the case for mental health data. On the other hand, OLS estimates of linear models evade the curse of dimensionality by imposing a rigid assumption of the underlying model, which reduces the expected error $\bar{e}$ to a linear function of the number of features L, given by Equation 2.13 [52].

$$\bar{e} = \sigma^2 \frac{L}{N} + \sigma^2 \tag{2.13}$$

where σ^2 is the variance of the error and N is the sample size.

It is worth noting here, that although least squares estimation reduces the effect of the dimensionality, compared to local methods, it is still desirable to increase the sample size and/or reduce the number of features to reduce the expected error.

2.5.4.2 Shrinkage Methods

In the presence of redundant features (especially linearly related features), the least squares estimates have low bias but large variance [52], which increases with the number of predictors. Consequently, shrinking or setting some of the coefficients to zero, reduces the variance and may improve the overall prediction accuracy. In addition, interpreting the prediction in terms of the predictors is easier with a fewer number of features that exhibit the strongest relationship to the outcome.

Ridge regression attempts to deal with multicollinearity by modifying the weights W^{ridge} to become [101];

$$W^{ridge} = \left(X^T X - \lambda I\right)^{-1} X^T Y. \tag{2.14}$$

By penalizing the computation of the coefficients, some coefficients are reduced compared to others. The amount of shrinkage depends on the value of λ.

A similar approach is Lasso regression, where the computation of the weights is also penalized. However, the penalty amount depends on the absolute value of the weights, unlike ridge regression, where the penalty is quadratic. This results in the lasso weights W^{lasso} not having a closed form for computation. On the other hand, the lasso has more opportunities to set coefficients to zero, and thus is better on the feature selection front [52].

2.5.5 Hidden Markov Model

HMMs [102] are a subset of statistical Markov models in which the system is modelled as a Markov chain with unobserved states. The uncertainty in the states is modelled through statistical distributions, which usually, but not necessarily, are Gaussian Mixture Models (GMMs) [103]. Although HMMs are powerful classifiers, they contain many parameters that need to be set manually, including: the number of states, the initial weights of the mixtures for GMMs and initial state transition probabilities, all of which affect performance [104]. HMMs can handle missing data given that the model is large enough to accommodate high variability in the inputs, but a large model is very complex to train. On the other hand, using feature selection along with HMM for classification encompasses training a new HMM for each patient, which is not feasible in real-time due to the memory-demanding nature of the training algorithm [105].

2.5.6 Independent Component Analysis

ICA [106] operates by locating independent axes within the signal space. ICA models data as a linear mixture of independent features using Maximum likelihood (ML) learning. Originally, Principal Component Analysis (PCA) was used to perform dimension reduction as a preprocessing step for ICA, which is not feasible to perform on data with missing features. In [107], the authors extend the use of Bayesian variational methods for dimension reduction with ICA, to handle missing data, by using the probability density estimate of the missing entries to fill in the missing val-

ues. Although this method is less biased than mean or regression imputation, it still fills in missing features with fabricated values. In addition, the training complexity is exponential in the number of features.

2.6 Predictions within GRiST

Initially, a tree model, eliciting expert-knowledge incorporated in the structure of GRiST, was developed to perform risk prediction. The idea was to propagate risk through the structure from leaf nodes to concept nodes in multiple levels up to the root node, which represented the risk. At each level the leaf nodes were combined using their Relative Influence (RI) to constitute a higher level concept as shown in Figure 2.7. Although the model provides impeccable explanation of the risk factors and perfectly resonates with expert-based models of risk, calculating RIs for M leaf nodes with N training examples involves solving N simultaneous equations with M variables, which is not feasible for large values of M and N [4].

Another approach involved developing a probabilistic graphical structure to model clinical expertise [108, 109]. However, the model required a lot of training and did not provide accurate predictions of clinical risk judgements [109].

The former approaches focused on developing a structure to accommodate the entire feature set available for risk prediction, relying on RIs or conditional probabilities to determine the importance of a feature to the prediction. However, finding RIs or probabilistic models that would work for all cases was challenging and complex. Contrarily, the Dynamic Feature Selection and Prediction (DFSP) [1], attempts to divide risk prediction into case-based problems by performing feature selection for each individual assessment record prior to risk prediction. Through feature selection, DFSP generates a specific set of cues that is particular to an assessment and uses only those cues to predict risk. The cues are selected based on their correlation to the risk while keeping a watch as to the amount of redundancy involved in the feature set. Risk prediction is performed using linear regression on the selected features. Though it had superior performance compared to other approaches when

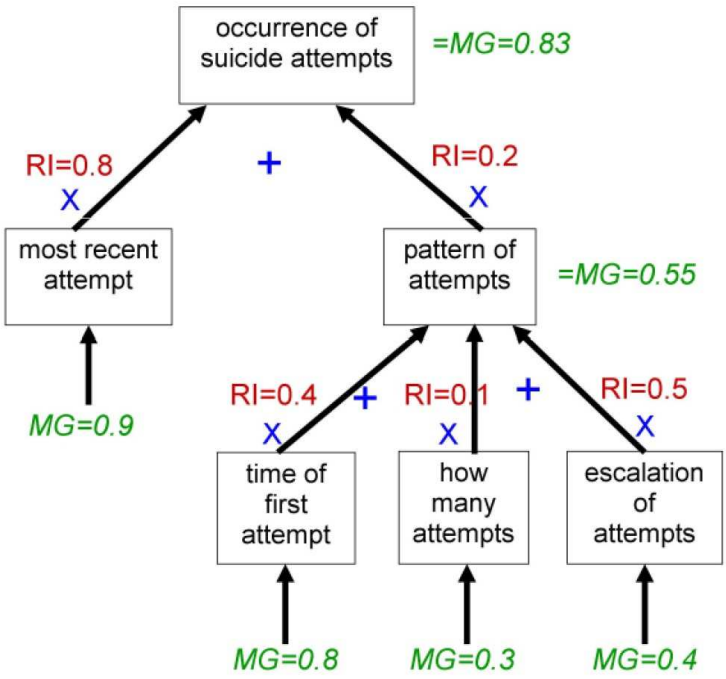

Figure 2.7: A portion of GRiST data for suicide risk showing how cue values are propagated using RIs, reproduced from [4]

tested within GRiST, some of the parameters in DFSP were required to be manually set. Section 2.7 explains the algorithm in details and highlights possible areas of improvement.

2.7 DFSP

DFSP [1] is an algorithm developed for risk assessment in mental health. DFSP divides the process into two distinct stages: feature selection and risk prediction. First, feature selection is performed based on a filter method with correlation with the DV as the decision metric. While the variables are added in order of their correlation with the DV, a threshold is applied to the mutual information between two IVs to reduce redundancy, such that; a variable is not added to the set, if its mutual information to the variables in the set exceeds that threshold. In [1] the algorithm is tested using different lengths for the feature set and it is concluded that 12 features are sufficient for risk prediction.

Second, for risk prediction, [1] tested several discriminative approaches including linear regression, MLR, GLM, DTs and random forests. While some performed better than others for specific classes, linear regression had the most consistent performance over all risk categories. To improve performance at the boundaries (highest and lowest risk categories), an adjustment (in 2.15) based on the distance between a prediction Y_{old} and the mean of the predictions $\bar{Y}$ is applied.

$$Y_{new} = Y_{old} + \frac{Y_{old} - \bar{Y}}{Y_{max} - Y_{min}} \tag{2.15}$$

where Y_{new} is the new prediction after adjustment and Y_{max} and Y_{min} are the upper and lower bounds of Y.

Although, DFSP outperforms other methods when applied to GRiST data, parameters such as the number of features and the mutual information threshold are manually chosen. In addition, DFSP provides no trade-off between correlation to the DV and interdependencies among features. Another shortcoming of DFSP is that it provides continuous unbounded predictions and does not perform classification or discretise the predictions and thus accuracy is measured solely on distance. Using a classification technique that minimises prediction error may improve performance. Finally, the adjustment provided by [1] is not justified and is empirically devised.

2.7.1 Updating Parameters

Whenever new data is available for training, the population changes, and hence, parameters used for feature selection and risk prediction will need to be updated. DFSP did not address the issue of autonomous update of parameters, but rather used predetermined values that were manually derived. The implementation of DFSP to different problems, other than the ones addressed in [1], will be hindered by the need to manually choose parameters for every new application.

2.7.2 Classification

Linear regression does not perform classification, and thus DFSP attributes predictions from the linear regression model to the nearest class. Using linear regression and then performing classification, reduces the classification problem to one dimension (the predictions), which means the problem may be treated as $K-1$ binary classification problems between K classes. From a classifier perspective, nearest neighbour in this context is equivalent to placing $K-1$ boundaries at $K-1$ midpoints between K classes. However, placing the decision threshold at the midpoint, will be optimum, only under the assumption that the probability distributions of the output of both classes have equal variance.

One of the assumptions of linear regression is homoscedastic errors [92]. Homoscedasticity implies that the variance of the error is constant for all prediction values. This assumption will be violated and heteroscedasticity will arise, if the number of training samples is not uniformly distributed throughout the range of the DV [110]. In mental health risk assessment, the size of each risk category varies widely which would cause severe heteroscedasticity in the prediction error. Since the variance of the error increases as the number of training samples decreases, high risk categories are affected the most by heteroscedasticity (since these usually contain the least amount of data). The variance is also affected by the shape of the probability distribution of the error, and thus the variance will be higher at the boundaries where the distributions are expected to be skewed. In such cases, placing decision thresholds at the midpoints between classes will be far from optimum.

2.8 Summary and Conclusions

Mental Health Risk Prediction faces various challenges, many of which arise from the nature of the data. First, parts of the data may not be collected, either because they do not apply to a particular case or they are just genuinely missing. Second,

the parts that are collected may or may not be crucial to a specific risk, as certain cues may be attributed to different risks and/or to risk management. Third, cues that may contribute to the risk for each patient may vary, even within the same risk domain. Finally, some cues are strongly correlated to one another, which causes redundancy in the data. A possible solution to the aforementioned problems is imputing the missing data. However, the sensitive nature of the domain and the massive amount of missing data, render imputing the missing data impractical and does not resolve issues like redundancy and relevance. Machine learning techniques, that can work around (dodge) missing data, will still use all the available data, of which some is not paramount to the risk. The irrelevant cues will add unnecessary complexity and dimensionality to a model, which makes the training process slower or even infeasible in some cases. Applying feature selection to each assessment individually is crucial to generating explainable results, as it produces a feature set that varies with different assessments.

Apart from the issues regarding the data, mental health risk prediction has to meet domain-specific needs. One of the most important considerations is the explainability of the risk prediction and the factors contributing to the risk. Selection and prediction problems could be jointly addressed using special types of Bayesian models that could be trained for high dimensional incomplete datasets [107]. Nevertheless, the complexity of such models would mean training and recognition would be too slow to be implemented in real time. Discriminative techniques, on the other hand, such as decision trees, SVM or NN have various limitations [88]. Decision trees have inherent feature selection capabilities and explainability and a computational class of the order $O(LNlogN)$ [111], but tend to get too dense to be comprehensible with high dimensional data. Contrarily, SVM and NN classifiers have no explanatory power and no feature selection capabilities, while their performance, though superior in some cases, is highly sensitive to the choice of model parameters. Diversely, regression models are easy to compute and are fully explainable. Multinomial logistic regression has built-in classification capabilities, but tends to underestimate the probability of rare events [98] and is more computationally exhaustive than linear regression. Linear regression analysis has the advantage of being simple, fast and explainable. However, the quality of the input feature set will influence re-

gression performance significantly. Collinearity between IVs, which is a result of redundancy in the feature set, will negatively affect the performance of regression models [53]. Furthermore, if linear regression is to be used for its competent speed and explainability, a feature selection algorithm has to be used, in conjunction, to select the most linearly correlated features to the DV, that are most independent of one another.

Since GRiST has a heterogeneous set of patients with high dimensional incomplete records, feature selection needs to be applied in a case-based manner. A feature selection mechanism that maximises the correlation of the selected features with the DV and simultaneously minimises redundancies within the feature set is paramount to the performance of a risk prediction algorithm. The complexity of the feature selection and risk prediction algorithms has to be minimised, to facilitate real time implementation.

On the other hand, shrinkage methods may be used for selection and prediction, conjunctively. However, these methods do not account for redundancies, which undermines the quality of the selected feature set. Moreover, the use of a single model to explain all cases would average out the variations among the different cases.

Chapter 3

Adaptive Feature Selection and Prediction

3.1 Rationale

In this chapter, a feature selection and risk prediction algorithm for mental health data is proposed. Feature selection is based on MRMR to simultaneously increase linear correlation to the risk being predicted while reducing the redundancy within the feature set. Linear correlation is used as the relevance metric as it maximises relevance, if linear regression is to be used for prediction. Linear regression is chosen for prediction, since it is the least computationally expensive algorithm when compared to other prediction and classification algorithms. Finally, a linear adjustment of the predictions is introduced and Unequal Variance Signal Detection (UVSD) is used to calculate decision boundaries used for classification, to account for heteroscedasticity that arises from the boundedness of the predictions and from inconsistencies in sample size among risk categories. The algorithm parameters are fully computable, which makes it possible to autonomously update the parameters as more data is collected and extend the applicability of the algorithm to various risk domains. The proposed Adaptive Feature Selection and Prediction (AFSP) algorithm, introduced in this chapter, is composed of two main phases: adaptive feature selection, and adjusted risk prediction and classification.

3.2 Adaptive Feature Selection

In feature selection, the selected feature set should adapt to include, not only factors that are relevant to the risk domain, in general, but also factors contributing to the risk in a particular assessment. The optimum feature set for each single assessment is one that truly maximises correlation with the dependent variable, while minimising the dependencies between the predictors, considering only the available data for the assessment in question. Given the large number of candidate features and no a priori knowledge of the optimum size of the feature set (that may even vary among assessments), the size of the search space is huge. Hence, the complexity of the problem is another factor to consider during feature selection. Through the three phases of feature selection presented in the following subsections, the complexity is being addressed as well as the fitness of the solution. In the first two stages, features are excluded based on either low correlation or GRiST ontology. The last step is applying feed forward feature selection based on MRMR Quotient, discussed in Section 2.4.2.

3.2.1 Correlation Threshold

An important aspect of feature selection is the relevance of the features to the prediction. In MRMR, the relative influence of the redundancy and relevance measures is difficult to adjust, which may lead to choosing features with low correlation to the DV, simply because they are independent to other variables. To ensure that the influence of correlation is not overpowered by mutual exclusivity of a candidate feature, cues whose correlation ρ to the predicted variable falls below a threshold ρ_{th}, are excluded from the candidate feature set. In addition, applying this threshold before MRMR contributes to reducing the search space and speeds up feature selection.

3.2.2 Filtering by the Ontology

Concept questions are gateways that open up a number of underlying questions. Though a concept may have a high correlation to the risk (on average), this does not dictate that all its underlying nodes have equal or high correlation to the risk. Children do not necessarily contribute equally to the influence of their parent node and thus do not contribute to the risk equally either. Whenever they are answered, children will contain more detail and information than their root node, since the presence of a value in a child necessarily dictates an MG of 1 for the filter, but a filter MG of 1 does not give any information about the child. Hence, when a child exists, its parent node is discarded from the candidate feature set, as the child holds more information including the parent's information. Not only does excluding parent nodes, for which the condition applies, reduce complexity by trimming down the search space, but it also improves the quality of the candidate set by removing redundant information before running MRMR.

3.2.3 Feed Forward MRMR Quotient

In order to trade off the amount of relevance and redundancy in a feature set, MRMR is used for feature selection. Redundancy R is measured through mutual information between features and relevance to the DV D is measured through linear correlation. Since the features are input to a linear regression model, linear correlation is the most appropriate relevance measure. On account of the redundancy and relevance measures not being comparable, a score based on the difference would require using weights to balance the influence of both terms, which would result in introducing new parameters to be computed, hence, using a score based on the quotient would overcome normalisation and scaling issues.

Finding the optimum feature set through MRMR encompasses trying out all possible combinations of features, which results in an exhaustive search of factorial complexity that may not be executed in real time, as discussed in 2. Instead a feed forward approach is chosen to reduce the complexity of the search to $O(M^L)$,

where L is the number of selected features and M is the number of candidate features. By adding features that are temporally optimum at each round, the search space is continuously reduced as the number of iterations increases.

3.2.4 Feature Selection Algorithm

At the beginning of feature selection, all eligible features are in the candidate set and the feature set is empty. In the first round of selection, the feature with the maximum correlation to the DV is selected first and removed from the candidate set. At each successive iteration the quotient scores Q of all remaining candidate features given the feature set are calculated. The feature with the minimum score is added to the feature set and removed from the candidate set. The selection moves forward only, thus features already added to the feature set in one round are never to be removed from the feature set in later rounds.

3.2.5 Parameters

Two variables are used in feature selection: the score Q and the overall score v. The score Q represents a ratio between: mutual information of a candidate feature with the existing feature set, and the candidate's correlation with the DV. Whereas the overall score v measures the fitness of the entire set. At each round of feature selection, the feature with the minimum score in Q is selected, until the overall score v reaches a predetermined threshold v_{th}, then feature selection stops.

First, linear correlation coefficients between all features and the DV are calculated, as the relevance measure, and mutual information values between all pairs of features are calculated, as the redundancy measure. To calculate the quotient score Q_m^i of a candidate feature Z_m at iteration i in Equation 3.1, the sum of the mutual information between the candidate feature Z_m and the features $X_1, X_2, ..., X_L$ already added in previous iterations is divided by the correlation D_m between that candidate feature and the DV.

$$Q_m^i = \sum_{l=1}^{L} \frac{R_{ml}}{D_m} \tag{3.1}$$

The scores measure the ratio of redundancy over relevance $\frac{R}{D}$, and thus the aim is to minimize the score. The same results of 3.1 may be obtained by Equation 3.3, if the individual scores matrix S is calculated and stored. Each element S_{ml} of S need not be calculated repeatedly at each iteration or for each patient and thus are calculated only once, off line, using Equation 3.2 and stored.

$$S_{ml} = \frac{R_{ml}}{D_m} \tag{3.2}$$

The quotient score Q_m^i of a candidate feature Z_m at iteration i is the sum of the individual scores between that feature and the features already added to the feature set, thus, during feature selection only the quotient scores in Equation 3.3 need to be computed in each round.

$$Q_m^i = \sum_{l=1}^{L} S_{ml} \tag{3.3}$$

Since the quotient scores increase incrementally only, previous iteration scores can be updated by individual scores at each round. Instead of performing L summations per candidate feature at each round, only one sum is needed per candidate feature to increment the score of that feature by S_{mL} using Equation 3.4, given that quotient scores of the previous iteration Q_m^{i-1} are stored.

$$Q_m^i = Q_m^{i-1} + S_{mL} \tag{3.4}$$

where $m = 1, 2, ..., M$, $l = 1, 2, ..., L$, and M and L are the lengths of the candidate feature set and the feature set, respectively.

3.2.5.1 Stopping Condition

In feed forward feature selection, the process will continue until a predetermined stopping condition is met. The stopping condition is usually based on prediction error and involves computing the error at each selection round. However, this wrapper approach is infeasible for high dimensional data with a large number of training samples. Instead, the stopping condition is determined off line and is periodically updated to accommodate new training data. At each iteration i, the sum of the quotient scores of previously added variables in Equation 3.5 is computed. The overall score v^i represents the MRMR score of the entire feature set, if this score exceeds a predetermined threshold v_{th}, the selection process stops.

$$v^i = \sum_{i=2}^{L} Q_i^i \tag{3.5}$$

3.2.5.2 Sample size

Figure 3.1: Sample size N against the number of features L at different values of the squared correlation coefficients ρ^2, reproduced from [5]

The number of samples available for training is an important factor in successfully predicting the output. Research conducted in [5] illustrated the relationship between the sample size and the number of features for multiple linear regression. Figure 3.1 shows that the sample size for a constant number of features varies at

different values of the correlation coefficients. However, for the worst case scenario of low correlation with the DV, the sample size is almost 100 times the number of features.

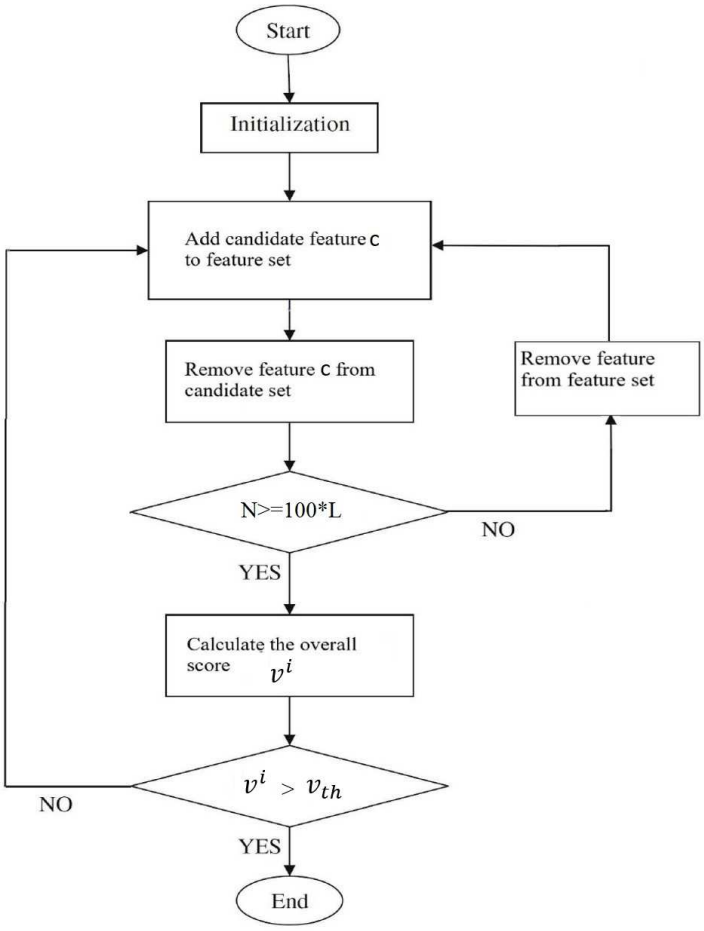

Figure 3.2: Feed-forward MRMR with a stopping condition v_{th}, N training samples and L added features

Before, permanently, adding a candidate to the feature set, the number of examples in the training set N with a complete feature set (including the candidate set) is calculated. If the sample size is less than 100 times the number of features L, the current candidate is not added to the feature set. However, feature selection will continue after the candidate in that round is removed from the candidate set, i.e. the sample size is not a stopping condition, since we move on to the next best feature that may have a larger sample associated with it. Figure 3.2 shows how the stopping condition and the sample size constraint are applied in conjunction

Algorithm 1: MRMRQ Feature selection

for every *testrecord* **do**
    Initialize *featset* to empty and *candset* to present features
    Initialize the score Q to empty, the overall score v to 0
    Set M to length of *candset* and l to 1 (l is the length of featset)
    Load S (S is the individual scores matrix)
    Add feature with maximum correlation to *featset*
    Remove feature with maximum correlation from *candset*
    while $v < v_{th}$ and $M > 0$ **do**
        for $m = 1$ to M **do**
            $Q_m = Q_m + S_{ml}$
        end for
        Sort Q in ascending order
        Sort *candset* with the same order as Q
        Add *candset[1]* to *featset*
        Remove *candset[1]* from *candset*
        $M = M - 1$ and $l = l + 1$
        Find a complete subset of *trainingrecords* for features in *featset*
        if $N < 100l$
        then Remove *featset[l]* from *featset* and $l = l - 1$
        else $v = v + Q_1$
        end if
    end while
end for

and Algorithm 1 details the feature selection process with all feed forward MRMRQ parameters incorporated.

3.3 Risk Prediction

Risk prediction is performed using linear regression for the previously mentioned reasons in Section 3.1. In this stage the features selected from the preceding stage for each individual assessment are used to predict the risk for that assessment. Training records that contain a matching feature set (i.e. with no missing entries) are compiled and used to calculate linear regression coefficients through OLS estimation.

The simplicity of OLS facilitates the real time feasibility of risk prediction. The linear regression weights directly link the feature set to the risk prediction and make the explanation of risk factors and their relative influence over the risk, straightforward. However, the nature of clinical risk judgements (the DV in this case) pose some constraints that may contradict the assumptions of linear regression; these are highlighted in the list below and fully addressed in the following subsections.

1. The DV is bounded, while the output of linear regression is unbounded.

2. Clinical risk judgements are discrete, whereas linear regression output is continuous.

3. The sample size is inconsistent among different risk categories which may lead to heteroscedasticity.

3.3.1 Heteroscedasticity

In mental health risk assessment, risk categories within a single risk domain will not have the same sample size, since high risk values are rare compared to lower values. The inconsistency in sample size leads to heteroscedasticity [112] and will cause the variance of the error to be different for each class. Furthermore, the boundedness in the DV, will cause a skewness in the distribution of the predictions close to the boundaries, as the model is fitted asymptotically [112]. Consequently, the variance of the error, and hence the error, will be higher at the boundaries and will decrease as the skewness fades away towards the center of the risk scale.

Since inconsistencies in the variance of the error violate the assumption of homoscedasticity adopted by linear regression analysis, methods to overcome heteroscedasticity are discussed in the following subsections.

3.3.1.1 Adjustment

A linear transformation, presented in 3.6, that scales and rotates the fitted line around the center, will improve performance at the boundaries [112]. The adjustment coefficients α and β in 3.6 are calculated by performing an auxiliary regression of the output prediction against the DV (as predictor) over the training set.

$$Y_{new} = \beta Y_{old} + \alpha \tag{3.6}$$

where Y_{new} is the new prediction and Y_{old} is the initial prediction before linear transformation.

3.3.1.2 Classification

Owing to the fact that clinical risk judgements are discrete, a mechanism for quantizing the continuous output from linear regression is needed. In other words, by using linear regression, the prediction and classification problems are separated and classification has to be performed based on the regression output. The classification problem in hand, may be treated as $K - 1$ binary classification problems between K classes, since the problem is solved in one dimension only.

The simplest way to solve binary classification problems is through equidistant decision boundaries, yet this only works if the output is homoscedastic. To account for the variation in the variance of the output, UVSD is used to calculate decision boundaries between any two classes.

3.4 Autonomous Parameter Update

It is essential that risk predictions in a mental health risk assessment system are provided in real time. However, other parameters utilised during feature selection and prediction need not be recalculated during individual assessments and may be

stored beforehand. These parameters are highly stable and can be calculated off line but they will need to be continuously updated as new data becomes available. The autonomous update of these parameters does not affect the speed of the real time tasks but enhances their performance through continuous updating of parameters.

Statistically derived parameters such as correlation and mutual information are recalculated by simply recomputing their equations over the new population. While decision boundaries are calculated based on UVSD and all thresholds are determined using a golden section search.

3.4.1 Decision Boundaries

Decision boundaries are calculated through solving quadratic equations whose coefficients are determined through Maximum Likelihood Estimation (MLE). To solve the classification problem, we choose the decision boundaries based on UVSD models [113]. The categories are represented by Gaussian distributions, hence, MLE of the means and variances is used [114]. The decision boundary λ_k between any two categories C_k and C_{k+1} that minimises the probability of error is obtained by solving 3.7 [115, 116].

$$\frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(\lambda_k - \mu_k)^2}{2\sigma_k^2}} = \frac{1}{\sqrt{2\pi\sigma_{k+1}^2}} e^{-\frac{(\lambda_k - \mu_{k+1})^2}{2\sigma_{k+1}^2}}, \forall k \tag{3.7}$$

where μ_k and μ_{k+1} are the means and σ_k^2 and σ_{k+1}^2 are the variances of classes C_k and C_{k+1}, respectively. Simplifying 3.7 yields 3.8 with a_2, a_1 and a_0 given by 3.9, 3.10 and 3.11 respectively. The threshold λ_k that minimizes the probability of error between any two neighbouring classes is obtained by solving 3.8.

$$a_2\lambda_k^2 + a_1\lambda_k + a_0 = 0, \tag{3.8}$$

$$a_2 = \sigma_k^2 - \sigma_{k+1}^2, \tag{3.9}$$

$$a_1 = 2(\mu_k \sigma_{k+1}^2 - \mu_{k+1} \sigma_k^2), \tag{3.10}$$

$$a_0 = \mu_{k+1}^2 \sigma_k^2 - \mu_k^2 \sigma_{k+1}^2 - 2\sigma_k^2 \sigma_{k+1}^2 \ln \frac{\sigma_k}{\sigma_{k+1}}. \tag{3.11}$$

where $k = 1, 2, ..., K-1$ and K is the number of classes.

3.4.2 Golden Section Search

The lack of a closed form to calculate the correlation threshold ρ_{th} and MRMR score threshold v_{th} calls for the use of a statistically guided search process. The aim of the search process is to find the threshold that minimises MSE of the regression prediction with respect to the clinical risk judgements before applying the decision boundaries. The problem of finding the correlation threshold ρ_{th} and the score threshold v_{th} may be broken down into two separate search problems, since the thresholds are applied in separate steps with no feedback.

A golden section search [117] is chosen to autonomously calculate the thresholds. The search process is performed for correlation first, since correlation threshold is applied before feature selection and no threshold is applied on the scores when performing predictions. When predictions are calculated for finding the score threshold v_{th}, on the other hand, the previously computed correlation threshold is used.

MSE is used as the control parameter when running golden section search. The MSE $f(t^j)$ of the predictions (when a threshold t is applied at iteration j) is calculated at the beginning, the value of t^j is, then, iteratively changed and MSE is recomputed at each round, which means that AFSP will be run for the entire dataset in each iteration of the search. The threshold t is changed such that the search space is

divided using the golden ratio $g = \frac{1+\sqrt{5}}{2}$ [118], where t^j is given by 3.12. Algorithm 2 illustrates the search process, where $[t_{max}, t_{min}]$ is the search range and is initially set such that t_{min} and t_{max} are the minimum and maximum of all possible values of t, respectively.

$$t^j = \frac{t_{max} + g t_{min}}{1 + g} \tag{3.12}$$

Algorithm 2: golden section search

Initialize $t^0 = t_{min}$ and $j = 1$
Calculate $f(t^0)$
$f(t^j) = f(t^0) + 1$
while $f(t^j) \neq f(t^{j-1})$ **do**
    Choose t^j in the range $[t^{j-1}, t_{max}]$
    Calculate $f(t^j)$
    If $f(t^j) < f(t^{j-1})$
        then $t^j = t^{j-1}$ and $t_{min} = t^j$
    else $t_{max} = t^j$ and $t^j = t_{min}$
    end if
end while

3.5 AFSP Summary

In this chapter, the AFSP algorithm was introduced. The algorithm is executed on 2 phases: feature selection, including preprocessing, and risk prediction. Feature selection commences with two preprocessing stages: applying a threshold on correlation and ontology-based filtering. The correlation threshold purpose is to reduce complexity and improve the quality of the candidate feature set. On the other hand, suppressing concept variables based on the ontology reduces redundancy in the candidate feature set. Feed forward feature selection is performed based on MRMR Quotient. A threshold over the entire MRMR score is used to stop feature selection

and a constraint over the sample size is applied to ensure sufficient training data exists whenever a feature is added.

The selected features are used to predict risk through linear regression analysis. The regression results are adjusted to account for heteroscedasticity. The decision boundaries used to categorise risk take into account the inconsistencies in the variance of the error. Finally, methods to autonomously update AFSP parameters off line are suggested.

Chapter 4

Application to Suicide Risk

4.1 Overview

Before we may implement AFSP for predicting suicide risk, the definition of what suicide risk encompasses has to be established. The concept of suicide risk assessment is not directly interpreted as the probability that a person dies by suicide. The person's intention, interpreted in suicide attempts indicates a degree of risk, even if the outcome is not fatal. As such para-suicide (attemped suicide) [17] is assessed within suicide risk, while fatal self-harm [119] is not the same as suicide, because the intention is absent. Self-harm is one of the predictors of suicide, but its risk is not evaluated within suicide risk assessment.

4.1.1 Scale Validation

During data collection within GRiST, the clinicians conducting assessments have placed a risk value in the range of 0 to 10. A value of 0 being the lowest risk (no risk) and a value of 10 being the maximum risk value. This results in suicide risk being evaluated on a scale of 0-10 in GRiST. Before we may implement an algorithm to estimate suicide risk based on the values in GRiST, the scale has to be validated, in order to establish the meaningfulness and repeatability of the responses. According

to [120] the establishment of prediction model in clinical practice (AFSP being one such model) has to go through four phases: development, validation, testing and implementation. The first three phases are crucial for implementing a statistical tool (phase four).

The development of AFSP has been discussed in Chapter 3. In our case, development is a complex construct, since AFSP relies on a structure from GRiST. While the development of GRiST (briefly discussed in Chapters 1 and 2) is not the objective of this thesis, as it has been developed prior to this work, validation of the suicide scale in GRiST is paramount to meaningful testing of the proposed algorithm.

Many measures of rater reliability exist [121], but their validity is subject to the domain. In mental health risk we cannot expect to have many clinicians assessing the exact same case or two clinicians dually assessing a large set of patients. Consequently, the method chosen to assess rater reliability should be able to deal with missing instances and multiple observers. Table 4.1 shows a comparison of various rater reliability measures in terms of how they deal with missing data, the number of observers, the interpretation of the output and the type of response [121].

From the comparison in Table 4.1, Krippendorf's α fulfils all the criteria and thus is chosen to compute inter-rater and intra-rater reliability for GRiST suicide scale. The next section details the computations of α in both cases.

Multiple assessments of the same patient by different clinicians, cannot be conducted at exactly the same time, thus the instances of each *unit* were not assessed at the same time. Since we are assessing the reliability of the suicide scale 0-10 and not the cue values, we have used cases for the same patient (the patient being the *unit*) when multiple assessors have placed a risk judgement of the same patient, at different times, and the cues have exactly the same values. The inter-rater reliability here is meant to measure how similar the judgements will be for the same cue values and the same person, to verify that the risk scale means the same to all the assessors.

Table 4.1: Comparison of rater reliability measures [6, 7, 8]

Criteria	Rater Reliability Measures			
	Percent agreement	Scott's π	Fleiss's K	Krippendorf's α
independent of the number of observers	✗	✗	✗	✓
invariant to permutation and selective participation	✗	✗	✓	✓
grounded in the distribution of the categories actually used by the observers	✗	✓	✓	✓
constitute a numerical scale with sensible reliability interpretations	✗	✓	✓	✓
work with ordinal data	✗	✗	✗	✓

4.1.1.1 Computing Krippendorf's α

The first step is to construct a reliability data matrix G, in 4.1 for p observers (clinicians) $O_1, O_2, ..., O_p$ for N *units* u, where N is the number of patients that are assessed by two or more different observers.

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1N} \\ g_{21} & g_{22} & \cdots & g_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ g_{p1} & g_{p2} & \cdots & g_{pN} \end{bmatrix} \tag{4.1}$$

The next step is to calculate the coincidences matrix B, in 4.3, where b_{ck} is given by:

$$b_{ck} = \sum_{u} \frac{n_{ck}}{n_u - 1}, \tag{4.2}$$

where n_{ck} is the number of $c-k$ pairs within a *unit* and n_u is the number of observations per *unit*. c and k take on different values of risk judgements, in our example there are 11 risk values 0-10 and this gives the range for both c and k.

$$B = \begin{bmatrix} b_{00} & b_{01} & \dots & b_{0k} \\ b_{10} & b_{11} & \dots & b_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ b_{c1} & b_{c2} & \dots & b_{ck} \end{bmatrix} \tag{4.3}$$

Because the data is ordinal, we need to calculate the ordinal metric differences δ_{ck}, given by:

$$\delta_{ck} = \sum_{h=c}^{h=k} n_h - \frac{n_c + n_k}{2}. \tag{4.4}$$

Finally, α is computed by Equation 4.5 for ordinal data. The value of α is in the range of 0 to 1, where 0 indicates complete disagreement and 1 indicates perfect agreement between observers.

$$\alpha = 1 - (N-1)\frac{\sum_{c}\sum_{k>c} b_{ck}\delta_{ck}}{\sum_{c}\sum_{k>c} n_c n_k \delta_{ck}} \tag{4.5}$$

For inter-rater reliability, data from different observers pertaining to the same *unit* is used, whereas for intra-rater reliability only data from the same observer in multiple assessments of the same *unit* is used.

4.1.2 Implementation

The goal of implementing AFSP in suicide risk prediction is to accurately predict the clinical risk judgements in real time. Implementation of AFSP within GRiST for suicide risk assessment is carried out in a series of steps, given below;

1. Feature selection parameters: relevance D, redundancy R and MRMRQ scores S; are computed and stored.

2. Golden section search is used to find the correlation threshold ρ_{th}.

3. Golden section search is used to find the score threshold v_{th}.

4. The adjustment parameters (α and β) are computed by an auxiliary regression.

5. Decision boundaries λ are calculated by applying UVSD.

6. Remove features from each pattern vector that are below the correlation threshold.

7. Remove all parent nodes from the candidate features if they have a descendant node present in the pattern vector.

8. Feature selection is performed over test data using feed forward MRMRQ.

9. OLS estimates of linear regression weights are calculated from training data.

10. Risk prediction is computed from the regression weights and the feature set.

11. Linear adjustment is applied to the initial prediction.

12. The prediction is classified using the decision boundaries from step 5.

Figure 4.1 shows steps 1 to 5 are performed to calculate the parameters required for AFSP, and thus are run off line on training data, while steps 6 to 12 simulate running feature selection, risk prediction and classification on test data. In this chapter the calculation of the parameters, required to apply AFSP to suicide risk, is discussed, while the results of running AFSP on test data are presented in chapter 5.

Training and testing is performed over 100,450 suicide risk assessments with clinical risk judgements from GRiST. 10-fold cross validation [122] is used to partition the data for training and testing, where the data is split into 10 equal sets and each set is used only once for testing while the remaining 9 sets are used for training as shown in Figure 4.2. The number of records per risk category N_{C_k} $(k = 1, 2, ..., 11)$

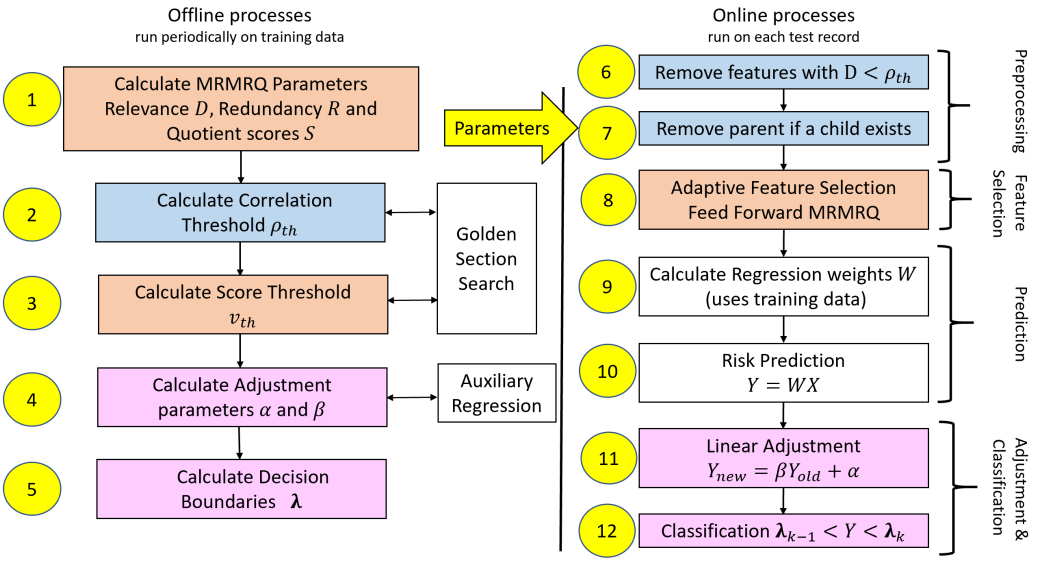

Figure 4.1: Steps of calculating parameters (1-5) and steps of applying AFSP to an assessment record(6-12); parameters in steps 1 to 5 are colour coded according to the stage of AFSP in which they are applied (Preprocessing, Feature Selection, Risk Prediction or Adjustment and Classification)

is divided equally among the 10 sets, to ensure that the risk categories have the same distribution in the training and testing sets as the whole data set.

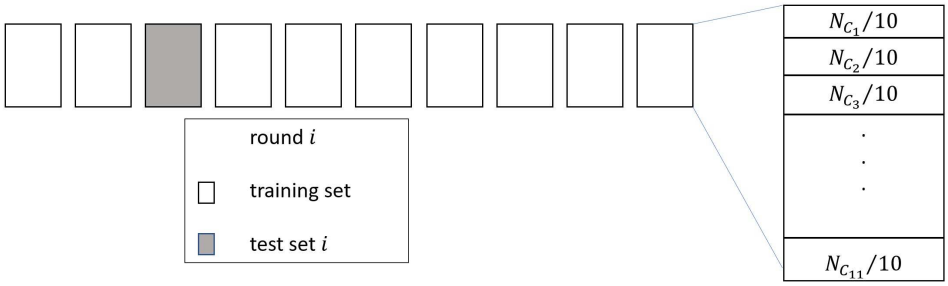

Figure 4.2: Data sets and the details of each set for round $i = 3$ of 10-fold cross validation

4.2 MRMRQ Parameters

In order to run MRMRQ for feature selection from suicide risk assessment data, the relevance and redundancy parameters D and R and the corresponding scores S need to be computed. The following subsections explain how the parameters are computed from suicide data and which parameters need to be stored for feature selection.

4.2.1 Relevance Parameter

The relevance parameter D is calculated using Pearson correlation coefficients. The correlation D_m between each IV X_m and the DV Y is calculated by 4.6. D constitutes the correlation vector of length M, where $M = 177$ and is the total number of features included in suicide risk assessment.

$$D_m = \frac{\overline{X_m Y} - \overline{X_m}\,\overline{Y}}{\sqrt{\overline{X_m^2} - \overline{X_m}^2}\sqrt{\overline{Y^2} - \overline{Y}^2}} \tag{4.6}$$

4.2.2 Redundancy Parameter

Mutual information is used to measure the redundancy R between two IVs. R is a matrix of size MXM $(177X177)$, where its elements R_{ml} are computed by 4.7 and denote the mutual information between any two IVs X_m and X_l.

$$R_{ml} = \sum_{x_m \in X_m} \sum_{x_l \in X_l} p(x_m, x_l) log\Big(\frac{p(x_m, x_l)}{p(x_m)p(x_l)}\Big) \tag{4.7}$$

4.2.2.1 Probability Distributions

The discrete probability distribution function of any IV X_m $(m = 1, 2, ..., 177)$ is computed by 4.8.

$$p(x_m) = \frac{N_{x_m}}{N_{X_m}} \tag{4.8}$$

where N_{x_m} is the number of occurrence of $X_m = x_m$ and N_{X_m} is the total number of instances of X_m. The joint probability of two IVs X_m and X_l is calculated by 4.9.

$$p(x_m, x_l) = \frac{N_{x_m \cap x_l}}{N_{X_m \cap X_l}} \tag{4.9}$$

where $N_{x_m \cap x_l}$ is the number of co-occurrence of $X_m = x_m$ and $X_l = x_l$ and $N_{X_m \cap X_l}$ is the total number of co-occurrence of X_m and X_l.

The probabilities and joint probabilities size will differ from one variable to another. For filter questions where there are only two possible values for X_m (0 or 1), $p(x_m)$ is a vector of length 2. Whereas for non-filter questions there are 11 possible values (0,0.1,0.2,...,1), and the length of $p(x_m)$ is 11. Possible sizes of the joint probability are 252, 2511, 1152 and 11511.

4.2.3 MRMR Quotient Scores

The IV individual MRMR scores matrix S used in feature selection is computed by scaling each row m of R by D_m, such that the individual score S_{ml} of feature X_m when feature X_l is a member of the feature set is given by 4.10.

$$S_{ml} = \frac{R_{ml}}{D_m} \tag{4.10}$$

4.2.3.1 Score Computations

In order to reduce the number of computations, the fact that $p(x_m, x_l) = p(x_l, x_m)$ is leveraged such that $R_{ml} = R_{lm}$, and thus one computation yields two elements of R except when $m = l$. The mutual information between a variable and itself is meaningless, and thus for $m = l$ mutual information need not be computed, and 1 is placed in the diagonal elements.

Since the elements on the diagonal of R are all set to 1, the diagonal of S will be the reciprocal of D. The values on the diagonal may be used for selecting the first variable in a feature set, as it is selected based on correlation only. The variable with the minimum diagonal value is selected in the first round of feature selection in Algorithm 3, since for the first iteration $Q = D^{-1}$. This reduces the memory

requirement of AFSP, since the correlations are not stored separately and only the individual scores matrix S needs to be stored.

Algorithm 3: Final feature selection algorithm

for every *testrecord* **do**
    Initialize *featset* to empty and *candset* to present features
    Initialize the score Q to empty, the overall score v to 0
    Set M to length of *candset* and l to 1 (l is the length of featset)
    Load S (S is the individual scores matrix)
    for $m = 1$ to M **do**
        $Q_m = S_{mm}$
    end for
    while $v < v_{th}$ and $M > 0$ **do**
        Sort Q in ascending order
        Sort *candset* with the same order as Q
        Add *candset[1]* to *featset*
        Remove *candset[1]* from *candset*
        $M = M - 1$ and $l = l + 1$
        Find a complete subset of *trainingrecords* for features in *featset*
        if $N < 100l$
        then Remove *featset[l]* from *featset* and $l = l - 1$
        else $v = v + Q_1$
        end if
        for $m = 1$ to M **do**
            $Q_m = Q_m + S_{ml}$
        end for
    end while
end for

4.3 Computing Thresholds

First, the correlation threshold and filter exclusion constraints are applied to the data, before feature selection commences. The correlation threshold is applied retrospectively, since it dictates which features may get into the candidate set, regardless of the available data.

Apart from MRMRQ parameters, a stopping condition v_{th} is required for a feed forward approach. In addition, the sample size constraint applied during feature

selection ensures that there is enough data to calculate regression weights for risk prediction.

4.3.1 Correlation Threshold

Figure 4.3 shows how golden section search is used to calculate the correlation threshold. When feature selection is performed during the search for the correlation threshold, the MRMR score threshold v_{th} is not applied and any feature may be included unless it evokes the sample size constraint. The aim of applying the correlation threshold is to suppress variables with very low correlation to the DV. Since it is a preprocessing step for feature selection, the correlation threshold is computed before the score threshold, as its value may affect the score threshold but not vise versa.

The correlation threshold ρ_{th} that gives the lowest MSE for suicide risk prediction is 0.15. Applying the threshold to the set of possible candidates reduces the number of all possible candidates from 177 to 76 features only. This means that the individual scores matrix S used during feature selection will need to include the coefficients corresponding to 76 features only and its size becomes $76X76$. The number of all possible candidates will change, if the correlation threshold changes when new data is available for recomputing the threshold. The length of the candidate feature set varies from one assessment to another because of missing features, but the maximum possible size is 76 (when all features are not missing), because of the correlation threshold.

4.3.2 Score Threshold

After the correlation threshold is calculated, another golden section search is performed to find the MRMR score threshold v_{th}. Since the correlation threshold is found beforehand, it is used during prediction when searching for the score thresh-

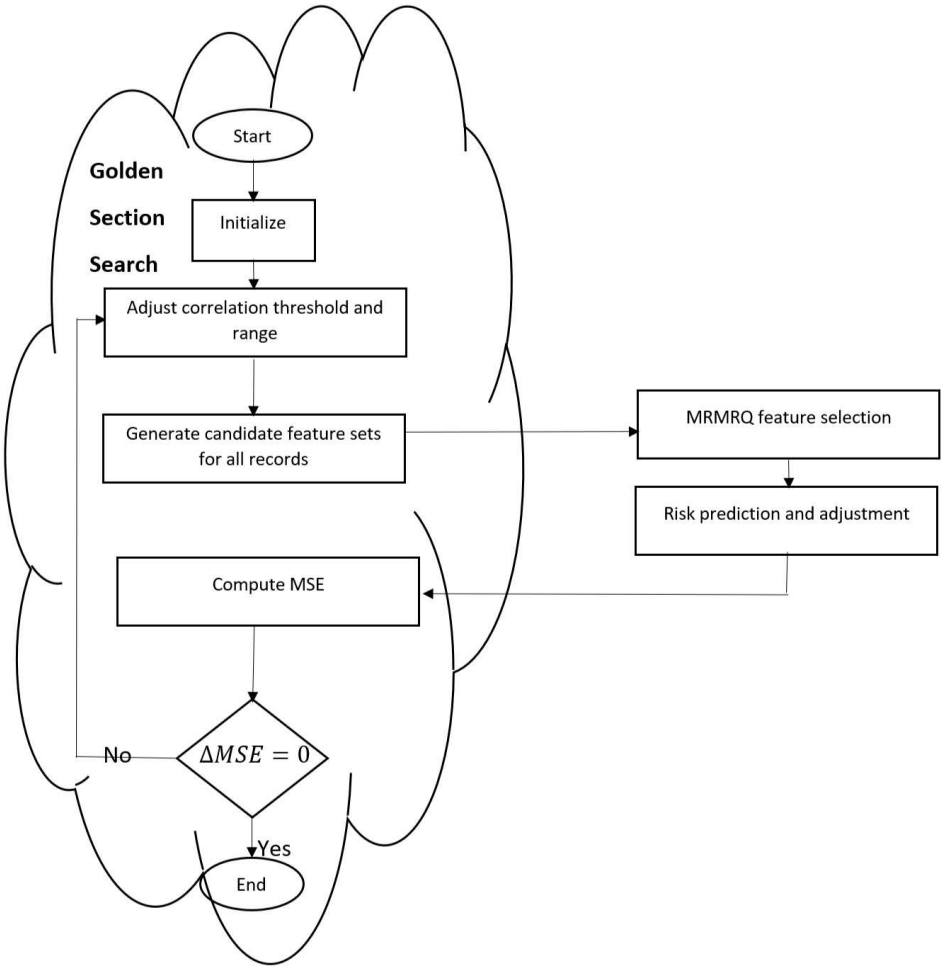

Figure 4.3: Using golden section search to compute correlation threshold ρ_{th}

old. Figure 4.4 explains how feature selection is applied during the search for the score threshold.

For suicide risk prediction the score threshold v_{th} that minimises MSE is 30. When the overall score v for a feature set reaches v_{th}, feature selection stops, as adding more features will not improve MSE and will add unnecessary complexity to the model.

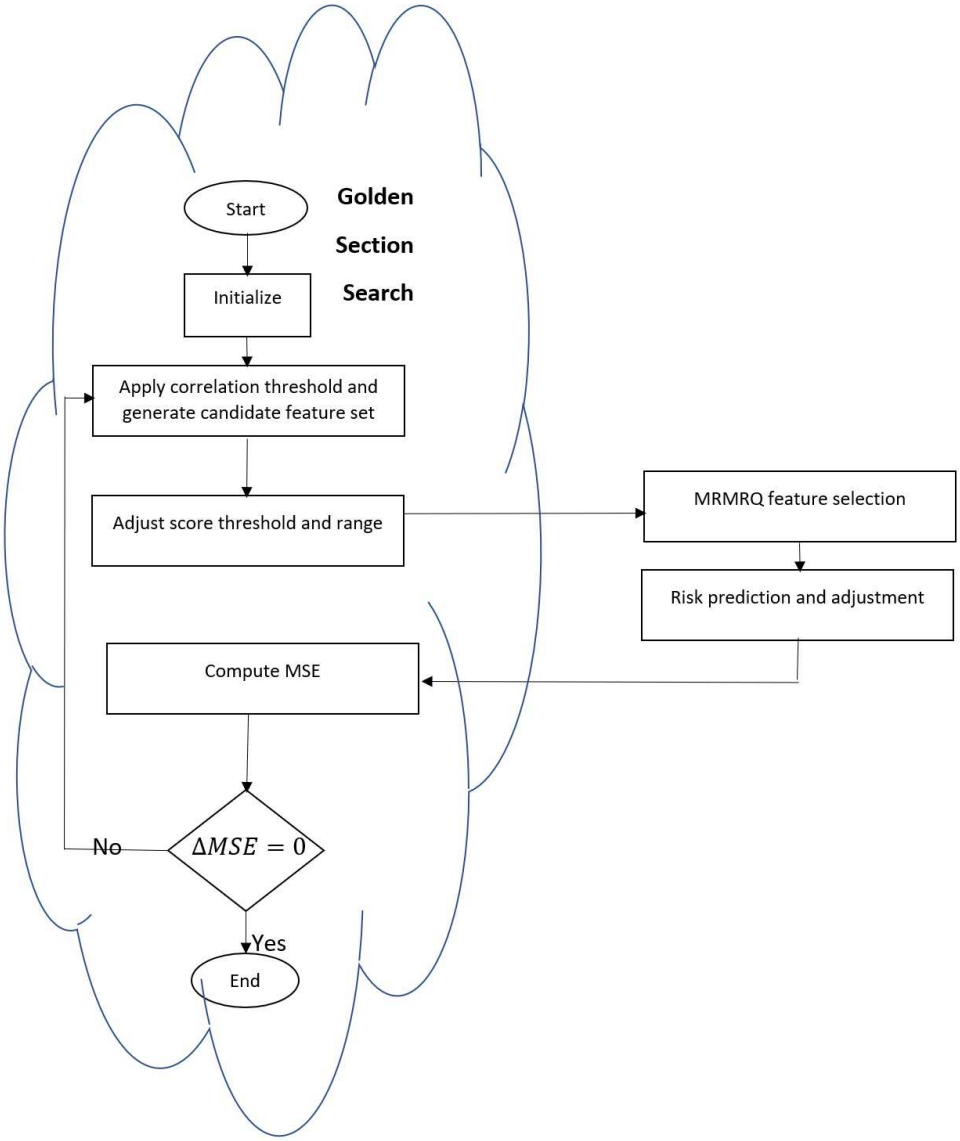

Figure 4.4: Using golden section search to compute score threshold v_{th}

4.3.3 Sample Size Constraint

The purpose of applying a constraint on the sample size, is to ensure that there is sufficient data, given the selected feature set, to train a prediction model. However, the sample size limit is not a stopping condition and will not stop feature selection altogether. Instead, the situation could arise where a variable that has been selected is actually dropped because of a violation of the minimum sample size, but feature selection continues and more variables are added. The variables added afterwards are worse from an MRMR point of view, since they came after the dropped variable

in the selection order, and thus potentially less influential variables could get into the set, after better variables have been dropped.

The deployment of the sample size constraint is essential, but it is equally important to measure its effect on feature selection. When feed forward MRMRQ feature selection is applied to suicide data in conjunction with the sample size constraint, only 7% of the records are negatively affected by the sample size constraint, while 93% of the times feature selection is performed, the selected features are not changed due to the sample size.

4.4 Suicide Risk Prediction

After feature selection, the sample for learning the regression model is obtained by finding all records that have a complete set of the selected features present. The weights for the model W are calculated using 4.11 [123] with clinical risk judgements Y as the DV. The IVs matrix X is $L+1$ columns where the first column is always 1 to calculate the intercept. The clinical risk judgements are scaled to $0, 0.1, 0.2, ..., 1$ corresponding to classes $0, 1, 2, ..., 10$.

$$W = (X^T X)^{-1} X^T Y \tag{4.11}$$

The products of the weights W and their associated variable values X are then used to predict the DV as shown in Equation 4.12.

$$Y = \sum_{l=0}^{L} W_l X_l \tag{4.12}$$

The predictions are classified into one of 11 (0-10) suicide risk categories C_k. The prediction is attributed to class C_k if the prediction Y satisfies 4.13.

$$\lambda_{k-1} < Y < \lambda_k \tag{4.13}$$

where λ_{k-1} is the boundary between classes C_{k-1} and C_k and λ_k is the boundary between classes C_k and C_{k+1}.

Note that for class C_1 there is no lower bound, since it is the lowest risk value, and thus the condition in 4.13 would become $Y < \lambda_1$ for C_1. On the other hand, the highest risk category C_{11} has no upper bound, and the condition in 4.13 becomes $Y > \lambda_{10}$ for C_{11}.

4.4.1 Testing for Heteroscedasticity

The uneven distribution of the number of samples N among suicide risk categories in Table 4.2, causes inconsistencies in the variance of the error in linear regression, since the variance decreases as the number of samples increases [123]. To test for heteroscedasticity, a Breusch-Pagan test [124] is performed over the predictions. The test confirmed that the homoscedasticity assumption is violated and that the results are heteroscedastic. In the following subsections the parameters used to adjust the results and classification parameters that account for heteroscedasticity are computed.

4.4.2 Adjustment Parameters

One cause of heteroscedasticity is the boundedness of the data, which skews the error distribution at the edges [112]. A possible solution for adjusting the variance at the edges is by performing an auxiliary regression. The aim is to adjust the predictions to match the clinical risk judgements. Hence, an auxiliary regression that considers the clinical risk judgements as the DV and the initial risk predictions Y_{old} as the IV, is devised.

Table 4.2: Sample size N_{C_k} against risk categories C_k

C_k	N_{C_k}
0	21,160
1	27,550
2	21,720
3	14,360
4	7,310
5	5,440
6	2,470
7	2,210
8	1,570
9	510
10	150
Total	10450

In this work, since 10-folds cross validation is used and the adjustment parameters are to be computed using training data, 10 different sets of values are obtained by regressing the predictions from training data over the clinical risk judgements. The average values, across 10 runs, of the auxiliary regression weights α and β in Equation 4.14 for suicide risk prediction are -0.02209302 and 1.16279069, respectively. These values are used to adjust the results, as they are more representative of the entire dataset, than values from a single round.

$$Y_{new} = \beta Y_{old} + \alpha \tag{4.14}$$

4.4.3 Classification Parameters

The decision boundaries are calculated based on UVSD, to account for the inconsistencies in the variance among classes. Figure 4.5 gives an example of the decision boundaries for class C_3 (risk value of 2) and their placement with respect to the conditional distributions $f_Y(y \mid C_2)$, $f_Y(y \mid C_3)$ and $f_Y(y \mid C_4)$ of the prediction Y given the class C_k. Table 4.3 shows the values of the boundaries λ_k between classes C_k and C_{k+1} for suicide risk predictions.

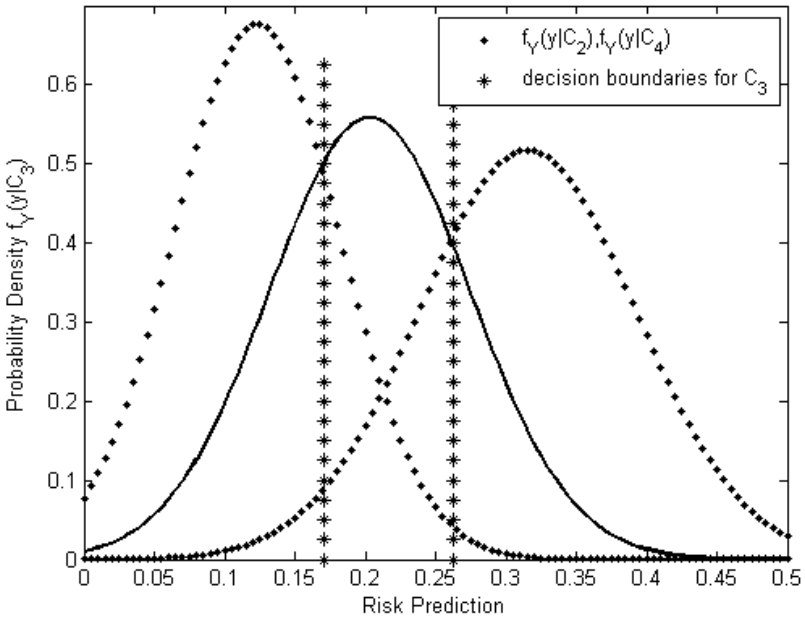

Figure 4.5: Classifier model for class C_3 showing three Gaussian distributions representing the conditional distributions of the predictions over the target class C_3 and the two neighbouring classes C_2 and C_4, along with the decision boundaries λ_2 and λ_3

Table 4.3: Classes C_k and C_{k+1} and the decision boundary λ_k separating them

C_k	C_{k+1}	λ_k
0	1	0.07101
1	2	0.17051
2	3	0.26310
3	4	0.34006
4	5	0.41909
5	6	0.49109
6	7	0.58020
7	8	0.66613
8	9	0.77010
9	10	0.91935

4.5 Summary

In this chapter, the computations and resulting values of off line parameters, required to run AFSP, have been discussed. An optimisation of the score matrix S is suggested to incorporate the correlation coefficients required for feature selection.

The calculation of multiple linear regression weights, performed on line for each assessment, is introduced. The results of the preprocessing steps, feature selection and risk prediction will be discussed in the following chapter, in terms of accuracy and speed.

Chapter 5

Results

5.1 Overview

In this chapter, the results of applying AFSP in suicide risk assessment within GRiST are presented. First, the results for suicide scale validation are presented, followed by risk prediction results. Predictive performance is measured through accuracy and shifted accuracy of predicting the clinical risk judgements. The shifted accuracy denotes predicting the clinical risk judgements to within ± 1 on the 0 to 10 risk judgement scale. Since 10-fold cross validation is used, the mean and standard deviation of the accuracy and shifted accuracy across the 10 runs are presented to indicate the stability of the results. Various feature selection methods, such as ReliefF, RReliefFF, MRMR variants and search strategies are compared to the proposed feed-forward MRMRQ approach. For prediction, linear regression is compared to Lasso regression and MLR.

In order to demonstrate the effect of the proposed preprocessing, adjustment and classification techniques, one parameter is added at a time. The following sections highlight the improvement in the results induced by introducing the correlation threshold, filtering by ontology, linear adjustment and UVSD boundaries.

The accuracy of the final version, with all constraints and parameters implemented, is compared to DFSP, since Nagy [1] shows that DFSP has superior performance in

suicide risk prediction compared to several other methods. In addition, DTs, Random Forests (RFs) and several combinations of selection and prediction techniques are compared to AFSP in terms of accuracy and stability under cross validation.

The last section details the speed and computational complexity of running all components of AFSP to ensure its capability to run in real time. All the simulations are performed using MATLAB™R2014a on an Intel®core™-i7 dual-core 2.7 GHz processor, with 4 logical processors, and 16 GB of memory.

5.2 Scale Validation

The GRiST suicide scale is internally validated by means of inter-rater and intra-rater reliability, given in the following Subsections.

5.2.1 Inter-Rater Reliability

In the sample of GRiST data used for the PhD there are over 12,000 patients with 104500 assessment records, but only cases that were assessed by multiple clinicians may be used to compute inter-rater reliability. The number of *units* u for which validation is applicable is 810, each has been assessed by two or more clinicians (observers) O, of 609 clinicians included in the measurement. The coincidence matrix B_{inter} is given by:

$$B_{inter} = \begin{bmatrix} 636.24 & 88.82 & 9.94 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 88.82 & 445.79 & 19.39 & 5 & 1 & 0 & 0 & 0 & 0 & 0 \\ 9.94 & 19.39 & 179.67 & 10 & 3 & 1 & 0 & 0 & 0 & 0 \\ 2 & 5 & 10 & 86.5 & 2.5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 3 & 2.5 & 52 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0.5 & 21.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 12 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 18 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} \tag{5.1}$$

Most of the non-zero values in B_{inter} are either on or near the diagonal, which indicates agreement between clinicians. The value of α_{inter} is found to be 0.8139.

5.2.2 Intra-Rater Reliability

The intra-rater reliability is evaluated only on cases that are assessed two or more times by the same clinician. The number of cases for which the condition applies is 458, assessed by 500 clinicians. The coincidence matrix B_{intra}, shown below, exhibits the same properties as B_{inter}, with significant values on or near the diagonal. The value of α_{intra} is equal to 0.9141 which indicates repeatability of the measurements.

$$B_{intra} = \begin{bmatrix} 291.33 & 15.83 & 0.83 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 15.83 & 329 & 4.17 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.83 & 4.17 & 145 & 3 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 3 & 3 & 63.5 & 1.5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0.5 & 37 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0.5 & 14.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 12 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} \tag{5.2}$$

Since the number of samples is non-uniform across the risk scale, with higher risk categories occurring less frequently than high risk ones, it is expected that the number of available instances of high risk *units* will be much lower as we move towards the higher end of the scale. This is manifested in the values of the coincidence matrices as these decrease moving towards the lower left corner.

5.3 Comparison of Feature Selection Techniques

For the feature selection component of the proposed algorithm, several viable alternatives are compared in terms of accuracy and stability. In order to neutralise the effect of the preprocessing and prediction components of AFSP on performance, all the preprocessing steps are applied in all cases, linear regression is used for prediction and classification is based on UVSD.

Figures 5.1 and 5.2 show a comparison between ReliefF, RReliefF and MRMRQ in terms of the average accuracy and shifted accuracy of the predictions. MRMRQ outperforms ReliefF and RReliefF across the entire risk scale. This may be owing to the fact that MRMRQ accounts for redundancy while Relief-based feature selection does not eliminate redundancy and only considers how well features separate

classes. Eliminating redundancies improves predictions based on regression, since it reduces co-linearity between features.

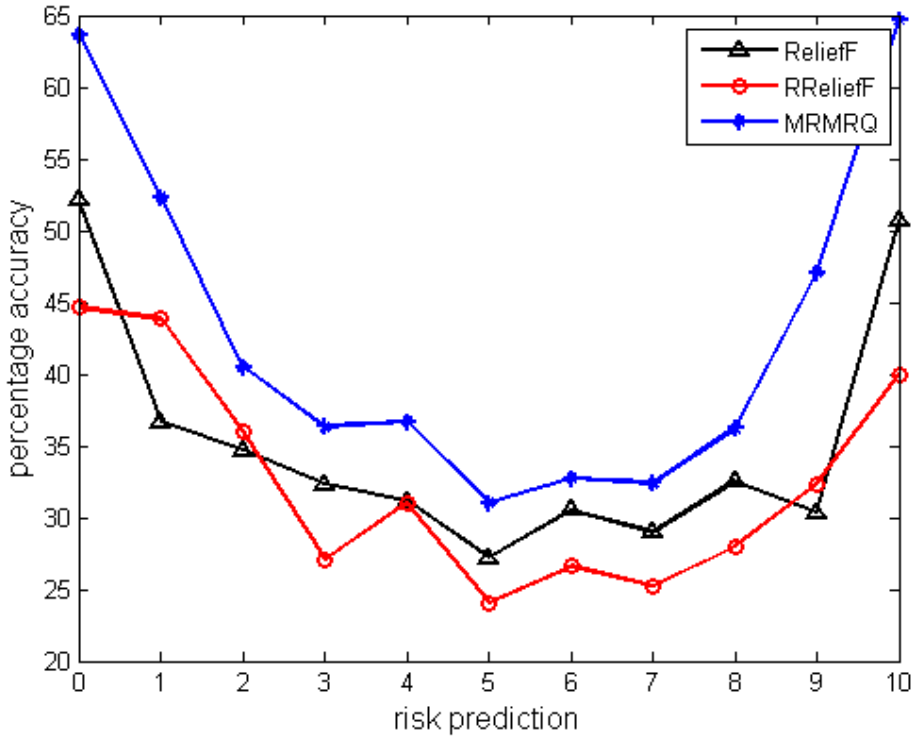

Figure 5.1: Percentage accuracy of AFSP across risk categories, with ReliefF, RReliefF and MRMRQ for feature selection

The stability of the results depends on the standard deviation of the results over the 10-folds, shown in Figure 5.3. The standard deviation of the accuracy of the 10 folds is expected to be higher for high risk categories, since these have the least amount of data. The standard deviation is lower for MRMRQ compared to both ReliefF and RReliefF across most risk categories, which indicates that MRMRQ performance is more stable than Relief.

Variants of MRMR, such as MRMR Difference (MRMRD) and MRMR with Normalized mutual information (MRMRN), suggested by Estévez *et. al* [73]; are also tested to choose the best alternative for feature selection.

Figures 5.4 and 5.5 show that MRMRD and MRMRQ outperform MRMRN over the entire scale, with MRMRQ having the highest accuracy and MRMRD as a close second.

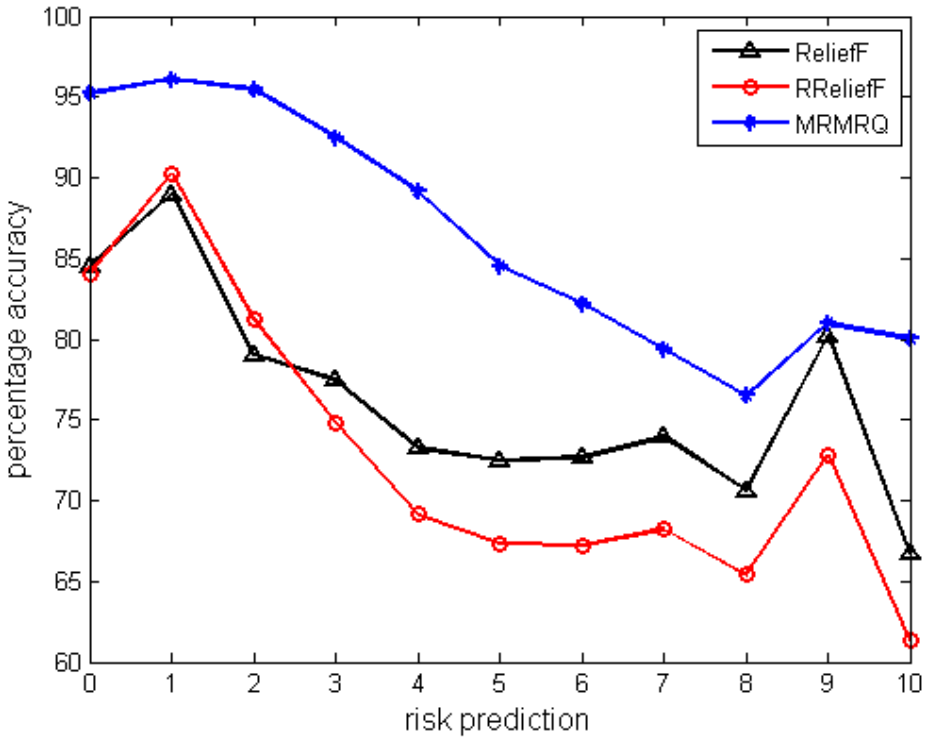

Figure 5.2: Percentage shifted accuracy of AFSP across risk categories, with ReliefF, RReliefF and MRMRQ for feature selection

The standard deviation of the accuracy, shown in Figure 5.6 indicates that the robustness of the results under cross validation is almost the same for the three MRMR variants.

5.3.1 Search Strategies

In this section, several sequential search methods are compared, such as forward selection, backward elimination, bidirectional search and floating selection. First, backward elimination is considered in comparison to forward selection. Figure 5.7 shows that forward selection has a higher average prediction accuracy. However, Figure 5.9 shows that backward elimination exhibits lower standard deviation at the edges, and thus its results are more robust than forward elimination.

A bidirectional search or a floating search may combine the merits from forward and backward techniques. Consequently, bidirectional search and FFS are tested. FFS

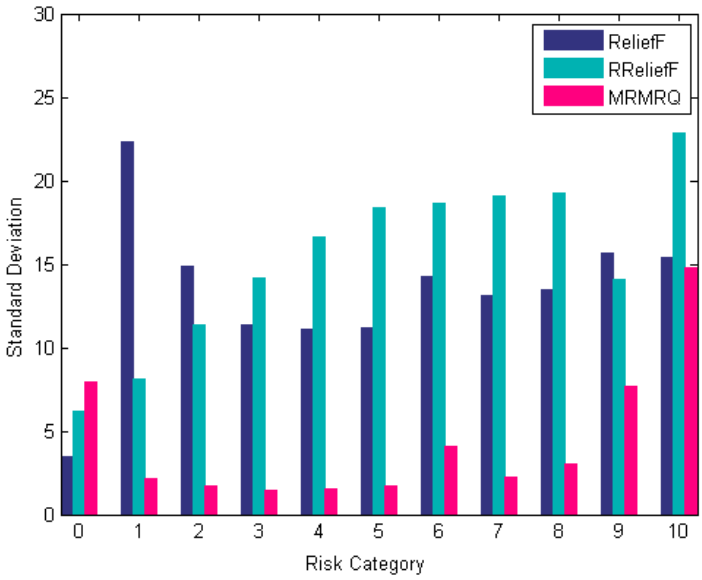

Figure 5.3: Standard deviation of the accuracy of AFSP across risk categories, with ReliefF, RReliefF and MRMRQ for feature selection

is chosen over FBE since forward selection outperforms backward elimination in terms of accuracy. Figures 5.10 and 5.11 compare the performance of bidirectional search and FSS, respectively, to forward selection. While the accuracy of the three methods is almost the same across all risk categories, the shifted accuracy is higher for forward selection, which indicates a lower prediction error. The stability of the three methods as indicated by the standard deviation in Figure 5.12 is almost the same.

5.4 Comparison of Prediction Techniques

For the prediction component of AFSP, three different approaches to regression are tested: Linear regression, Multinomial Logistic Regression (MLR) and Lasso regression. Linear regression is chosen since it is the least computationally intensive method, logistic regression eliminates the need for classification and Lasso re-

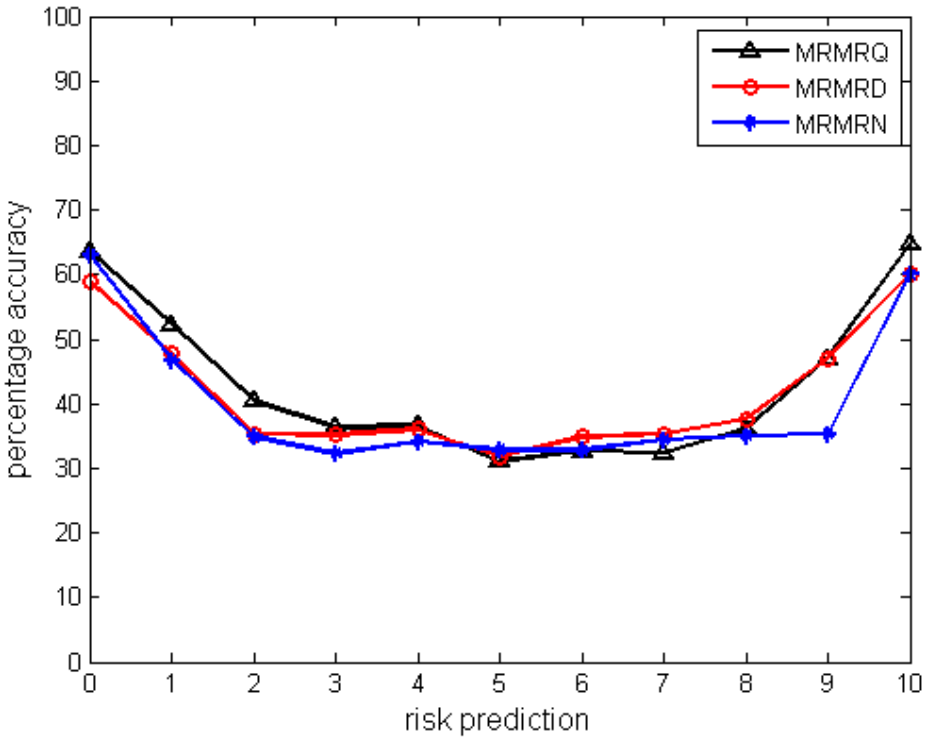

Figure 5.4: Percentage accuracy of AFSP across risk categories, with MRMRQ, MRMRD and MRMRN for feature selection

gression has inherent feature selection capabilities, which eliminates the need for computing regression coefficients online.

Figure 5.13 shows that linear regression outperforms MLR and the Lasso across the entire risk scale, Lasso regression comes as a close second, whereas logistic regression has the worst performance.

The stability of the results over cross validation is indicated by the standard deviation in Figure 5.14. Logistic regression is highly affected by sparsity, compared to the Lasso and linear regression. Consequently, the results of logistic regression are less robust for high risk categories, while Lasso and linear regression are almost equally stable.

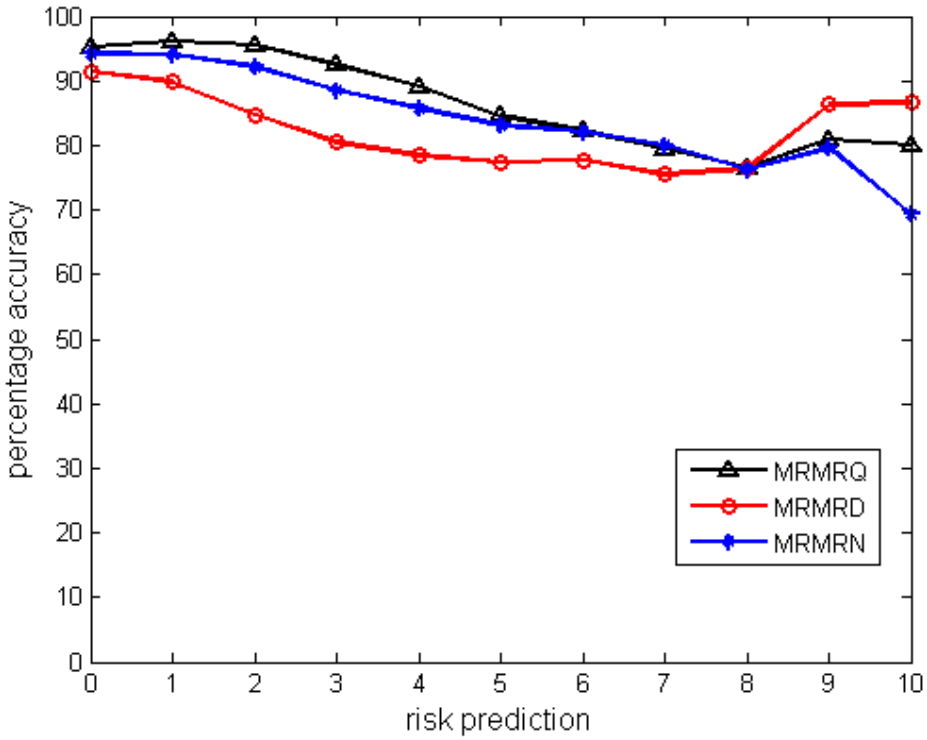

Figure 5.5: Percentage shifted accuracy of AFSP across risk categories, with MRMRQ, MRMRD and MRMRN for feature selection

5.5 Applying Correlation Threshold

The correlation threshold is applied in isolation, while the concept exclusion criteria, adjustment and UVSD classification are not applied. Table 5.1 shows that the prediction accuracy for all risk categories with and without applying the correlation threshold is almost the same. However, the mean length of the selected feature set decreases from 16.43 to 13.99, when the correlation threshold is applied, which means that the same performance may be achieved with a smaller feature set. A lower number of features reduces the computational complexity of risk prediction. Moreover, the average candidate feature set size drops from 64 to 28 cues and the maximum candidate feature set size drops from 177 to 76 features, which speeds up the feature selection process. The improvement in speed is discussed in section 5.10.

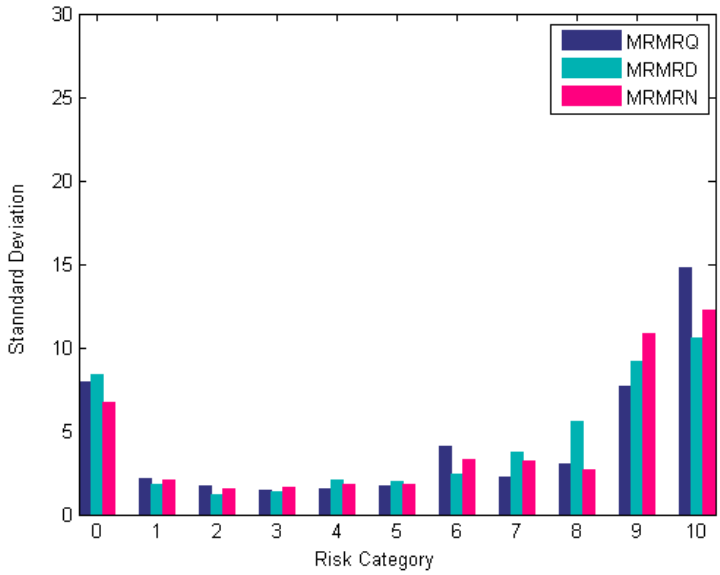

Figure 5.6: Standard deviation of the accuracy of AFSP across risk categories, with MRMRQ, MRMRD and MRMRN for feature selection

Table 5.1: Classifier statistics showing the percentage accuracy and shifted accuracy (S. accuracy) at each risk level with and without the correlation threshold

Risk	without correlation threshold		with correlation threshold	
	Accuracy	S. accuracy	Accuracy	S. accuracy
0	35.75	89.47	35.91	89.91
1	54.64	92.12	55.72	93.30
2	39.59	89.32	40.46	90.65
3	35.17	82.61	36.73	83.62
4	31.28	76.62	31.16	76.81
5	25.94	68.73	24.36	66.07
6	24.04	64.21	20.56	59.19
7	21.08	59.00	18.05	54.61
8	15.60	52.86	13.88	48.40
9	8.62	43.33	9.41	42.54
10	0.66	26.66	0.66	34.66

5.6 Applying Concept Exclusion

The concept-descendant exclusion constraint is applied after the correlation threshold, while the linear adjustment and UVSD are not applied. Table 5.2 shows that

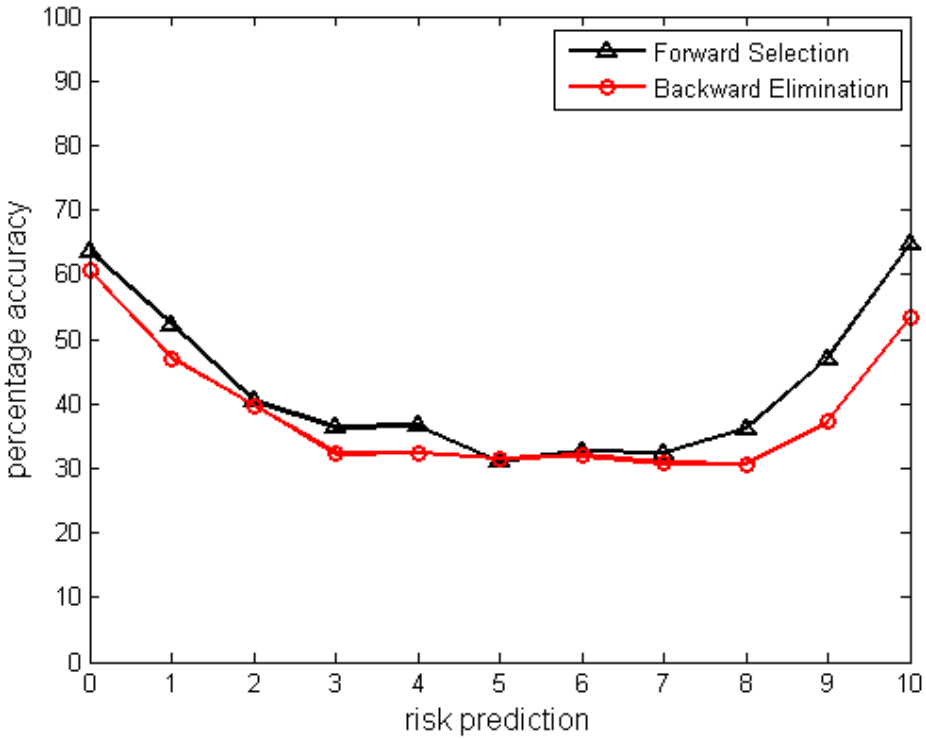

Figure 5.7: Percentage accuracy of AFSP across risk categories, with forward selection and backward elimination

applying the concept exclusion condition significantly improves the results of high risk categories but does not affect low risk prediction accuracy. Although the number of high risk assessments is comparatively low, it is particularly important to predict these cases accurately, since high risk cases are more critical and need attention and management. The improvement in the accuracy for high risk patients is mainly credited to the fact that high risk patients have more filter questions answered as "Yes" (since "Yes" indicates an area of concern), and thus more answers in leaf nodes, compared to low risk patients.

Filtering by ontology does reduce the size of the candidate feature set, but since it is performed for each case individually before MRMR, the reduction in computational complexity of feature selection is counteracted by the complexity of searching for filters and children through an assessment record, and therefore it does not a significant reduction in complexity.

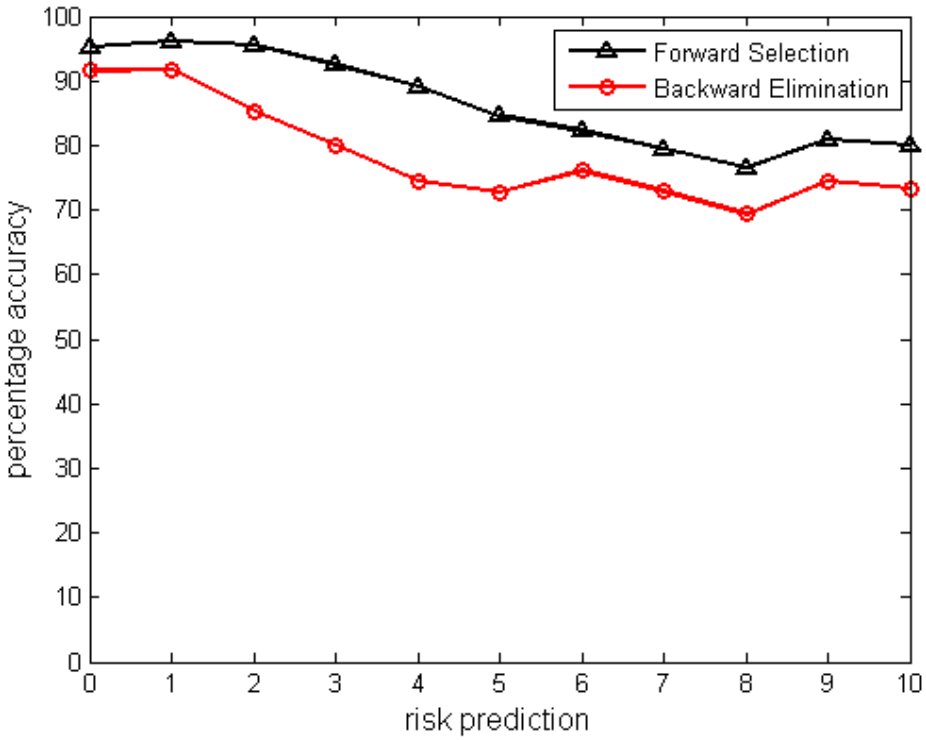

Figure 5.8: Percentage shifted accuracy of AFSP across risk categories, with forward selection and backward elimination

Table 5.2: Classifier statistics showing the percentage accuracy and shifted accuracy (S. accuracy) at each risk level with and without concept exclusion

Risk	without concept exclusion		with concept exclusion	
	Accuracy	S. accuracy	Accuracy	S. accuracy
0	35.91	89.91	35.10	90.51
1	55.72	93.30	54.96	92.99
2	40.46	90.65	41.23	91.13
3	36.73	83.62	35.89	85.37
4	31.16	76.81	31.25	76.90
5	24.36	66.07	23.78	70.16
6	20.56	59.19	19.99	63.68
7	18.05	54.61	20.36	54.48
8	13.88	48.40	15.56	53.11
9	9.41	42.54	13.54	53.54
10	0.67	34.67	14.67	55.33

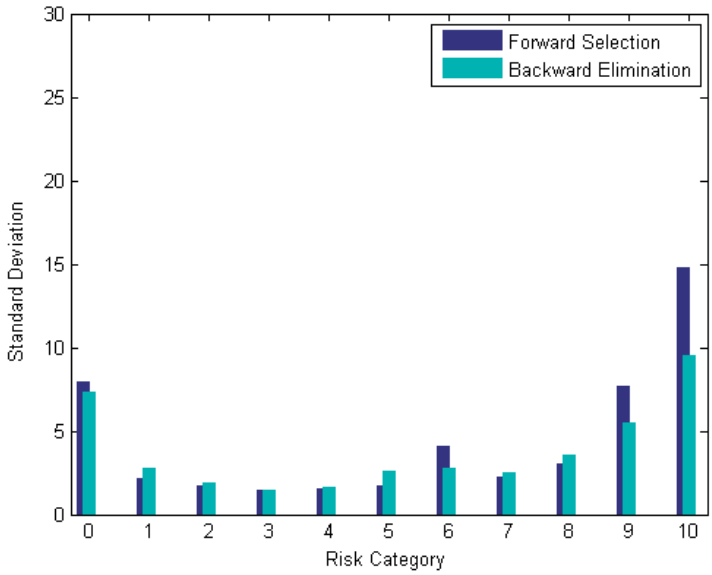

Figure 5.9: Standard deviation of the accuracy of AFSP across risk categories, with forward selection and backward elimination

5.7 Linear Adjustment Results

The linear adjustment is applied after risk prediction using linear regression, to account for heteroscedasticity, especially at the boundaries. Figure 5.15 shows that the mean risk predictions are a better match to clinical risk judgements after the adjustment. The mean absolute error in the predictions in Figure 5.16 has significantly decreased for risk categories 0, 8, 9 and 10, but has increased only slightly for risk values 1,2 and 3. Moreover, the errors are more consistent after the adjustment, since heteroscedasticity has decreased.

To highlight the effect of the adjustment on different risk categories, the distribution of risk predictions against clinical risk judgements is illustrated for the minimum (0), median (5) and maximum (10) risk values in Figures 5.17, 5.18 and 5.19, respectively. The distribution of the risk predictions is almost the same before and after the adjustment for a risk of 5, but for the boundaries (0 and 10) the center of the distribution is shifted towards the true value of the class. The mean risk predic-

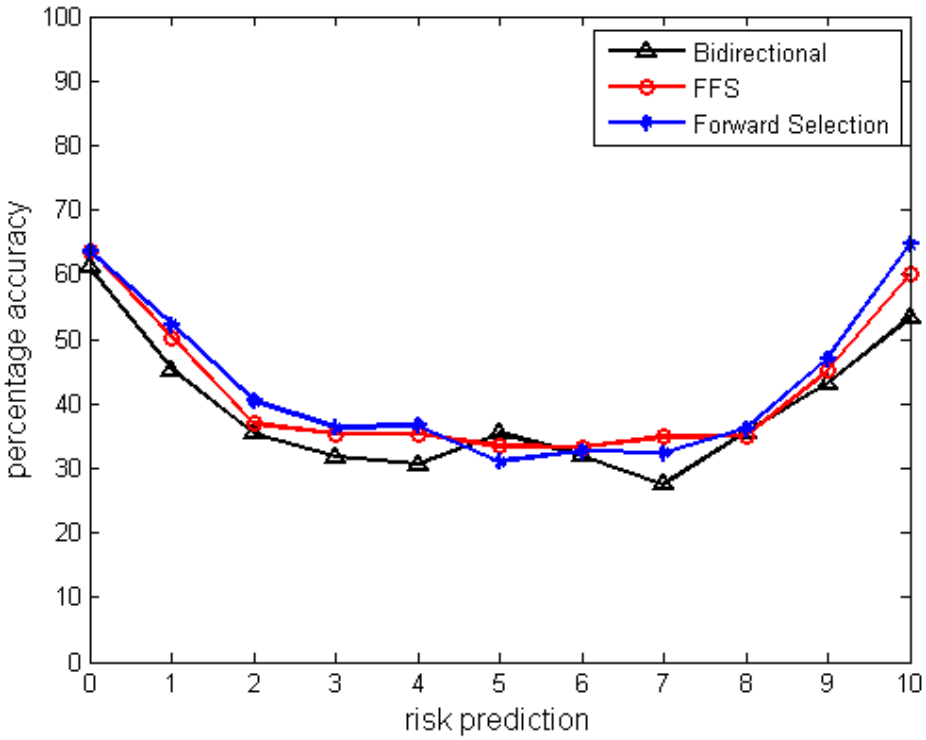

Figure 5.10: Percentage accuracy of AFSP across risk categories, with forward selection, bidirectional and FFS

tion for class 0 decreases from 0.0898 to 0.0823, while the mean of the predictions for class 10 increases from 0.8614 to 0.9796, when the adjustment is applied.

Table 5.3: Classifier statistics showing the percentage accuracy and shifted accuracy (S. accuracy) at each risk level before and after adjustment

Risk	before adjustment		after adjustment	
	Accuracy	S. accuracy	Accuracy	S. accuracy
0	35.10	90.51	47.04	90.21
1	54.96	92.99	52.83	92.22
2	41.23	91.13	42.89	87.83
3	35.89	85.37	40.65	79.04
4	31.25	76.90	39.19	74.94
5	23.78	70.16	30.95	71.86
6	19.99	63.68	26.23	71.54
7	20.36	54.48	24.80	68.72
8	15.56	53.11	26.69	76.08
9	13.54	53.54	44.90	76.08
10	14.67	55.33	58.67	76.00

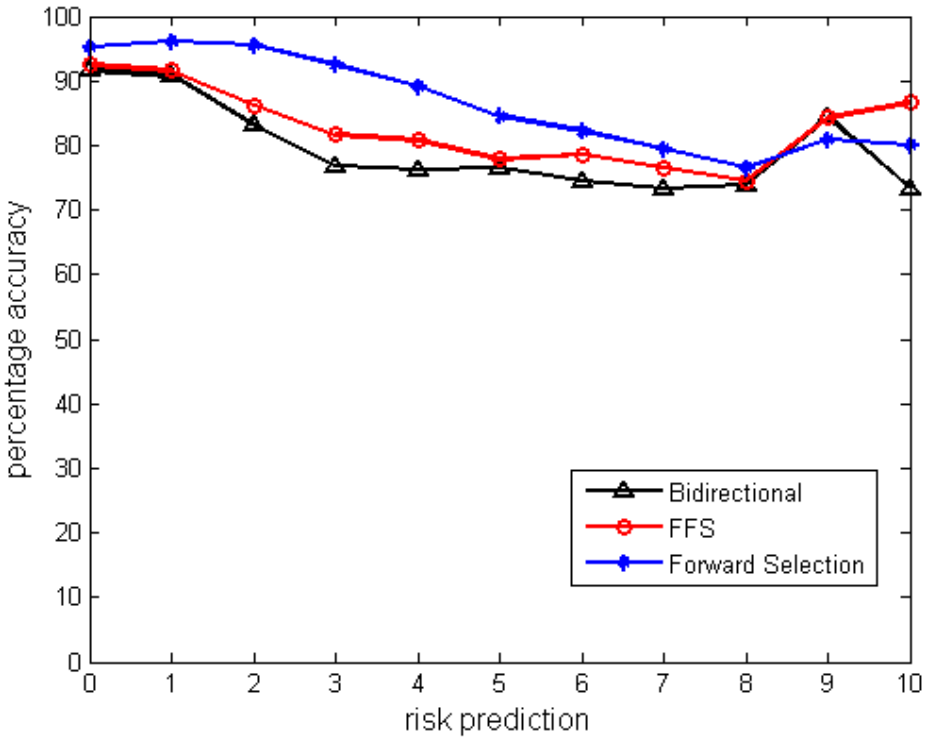

Figure 5.11: Percentage shifted accuracy of AFSP across risk categories, with forward selection, bidirectional and FFS

Table 5.3 shows that the accuracy of the prediction is significantly higher for high risk categories after adjustment, while low risk categories, except for 0, are not affected. The shifted accuracy, on the other hand, is negatively affected for classes 2-5, but this improves shifted accuracy for classes 6-10, hence, the performance is more consistent after adjustment. The correlation threshold and concept exclusion are applied and equidistant boundaries are used for classification in both cases (before and after adjustment) displayed in Table 5.3.

5.8 UVSD Results

Classification is the last phase of AFSP and is achieved through the deployment of UVSD boundaries. Figure 5.20 shows that the mean absolute error is more consistent when the proposed decision thresholds are used, compared to placing thresholds at the midpoint between two classes. Table 5.4 highlights the improvement in

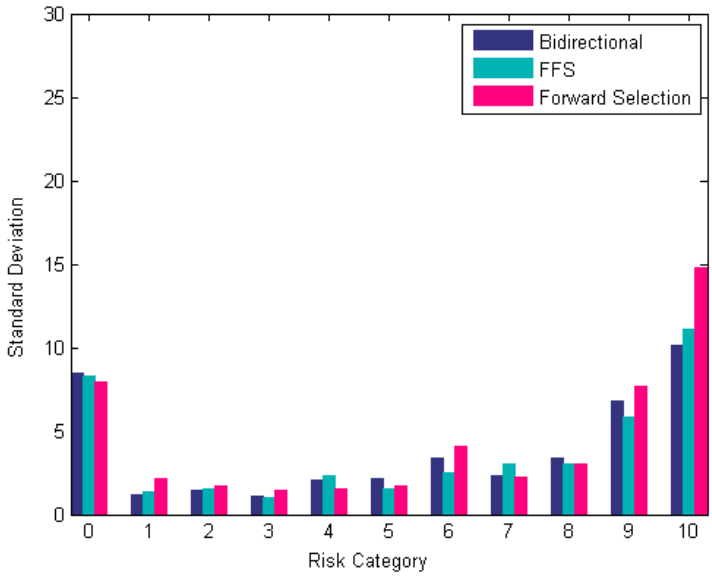

Figure 5.12: Standard deviation of the accuracy of AFSP across risk categories, with forward selection, bidirectional and FFS

performance for classes 0 and 6-10 when UVSD is used. The performance is the same or worse for risk values 1-4, but the shifted accuracy is better with UVSD over all risk categories. This is mainly because the magnitude of the error is lower, since the variance is more consistent after the adjustment, but the number of errors may slightly increase for some classes. The increase in the number of errors for certain classes is accompanied by a decrease for others, since the boundaries are chosen to maximise benefit for two classes at a time.

5.9 Comparison to Alternative Approaches

The accuracy and shifted accuracy of AFSP is compared to several other approaches, including DFSP, since the latter offers superior performance over other methods such as Decision Trees, Random Forests and correlation based feature selection followed by linear regression [1], when applied to GRiST data. DFSP is simulated over the same data, using the same hardware and with the same 10-fold partition-

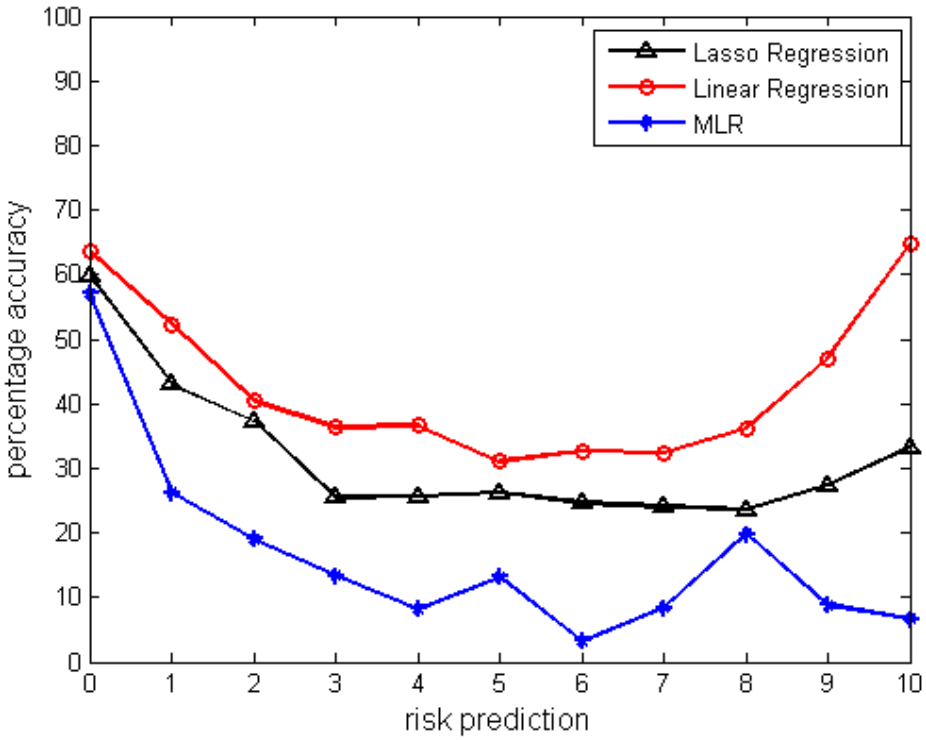

Figure 5.13: Percentage accuracy of AFSP across risk categories, with linear regression and MLR compared to a single Lasso regression model

Table 5.4: Classifier statistics showing the percentage accuracy and shifted accuracy (S. accuracy) at each risk level with equidistant boundaries and UVSD boundaries

Risk	Equidistant boundaries		UVSD boundaries	
	Accuracy	S. accuracy	Accuracy	S. accuracy
0	47.04	90.21	63.67	95.24
1	52.83	92.22	52.29	96.10
2	42.89	87.83	40.45	95.47
3	40.65	79.04	36.33	92.51
4	39.19	74.94	36.66	89.15
5	30.95	71.86	31.00	84.51
6	26.23	71.54	32.71	82.19
7	24.80	68.72	32.40	79.41
8	26.69	76.08	36.21	76.50
9	44.90	76.08	47.06	80.98
10	58.67	76.00	64.67	80.00

ing. Table 5.5 shows that the accuracy and shifted accuracy of AFSP is higher than DFSP across all risk classes. However, the improvement is not constant across all risk values, since the consistency of AFSP across the risk scale is better than DFSP.

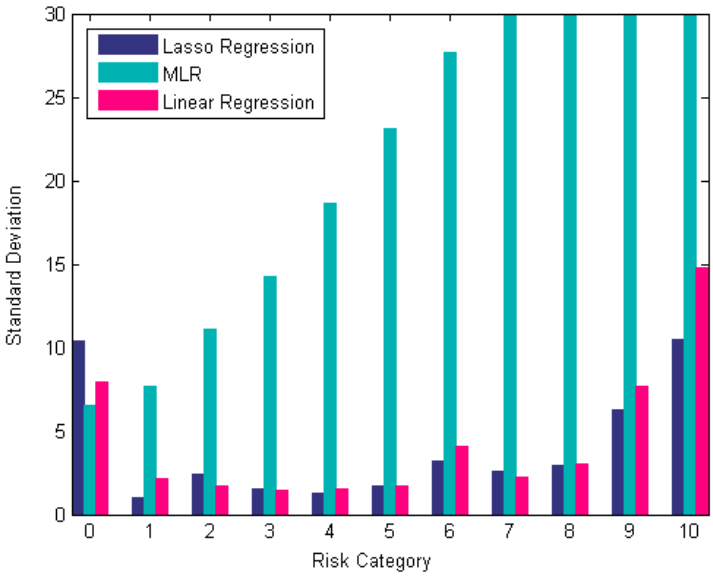

Figure 5.14: Standard deviation of the accuracy of AFSP across risk categories, with linear regression, MLR and a single Lasso regression model

Table 5.5: Classifier statistics showing the percentage accuracy (Acc) and shifted accuracy (S.Acc) of predictions based on: Decision Trees (DT), Random Forest (RF), ReliefF and Multinomial Logistic Regression (ReliefF), Correlation-based Feature Selection and Linear Regression (CFS+LR), and Minimum Redundancy Maximum Relevance Difference with Floating Forward Selection and linear regression (MRMRD); compared to DFSP and AFSP

C	DT		RF		ReliefF		MRMRD		DFSP		AFSP	
	Acc	S.Acc	Acc	S.Acc	Acc	S.Acc	Acc	S.Acc	Acc	S.Acc	Acc	S.Acc
0	69	95	63	96	14	90	64	95	57	91	64	95
1	47	96	60	98	67	94	50	94	46	92	52	96
2	36	85	45	91	27	84	37	84	35	84	40	95
3	41	76	34	78	30	63	35	68	33	78	36	93
4	17	72	14	62	12	66	35	66	32	74	37	89
5	29	45	22	36	33	46	33	36	28	71	31	85
6	9	58	8	45	0	59	33	51	27	69	33	82
7	34	54	31	43	27	45	35	36	29	71	32	79
8	30	67	27	57	42	64	35	46	31	69	36	77
9	24	63	9	42	7	72	45	34	37	75	47	81
10	32	53	6	9	2	31	60	9	44	64	65	80

DTs and RFs have the least consistent performance across risk values. ReliefF does not perform well for high risk patients because these are likely to have more data

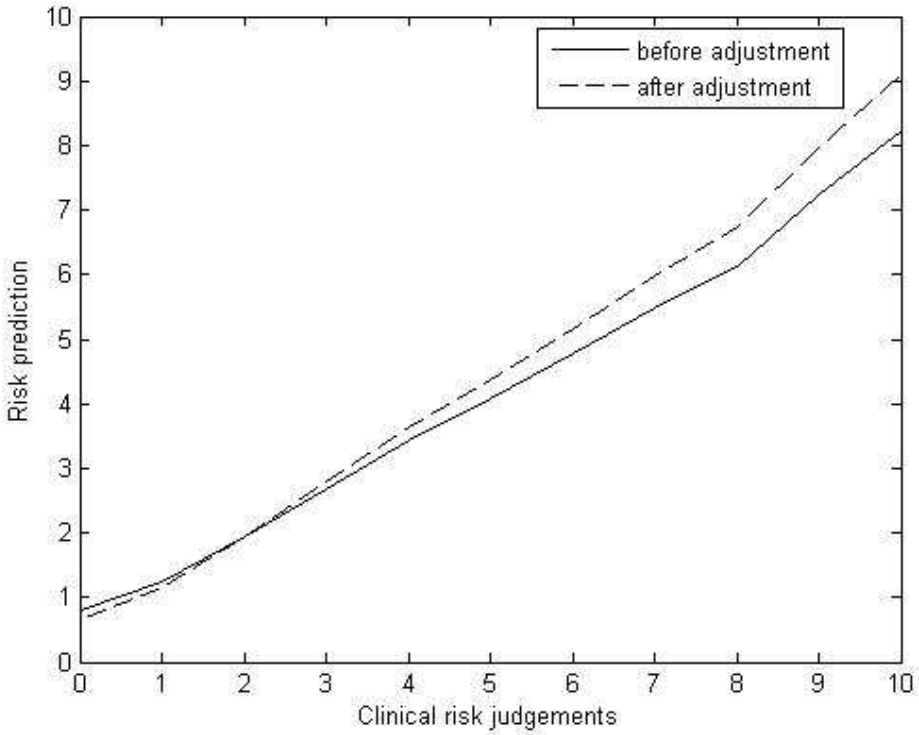

Figure 5.15: Mean risk predictions against the clinical risk judgements before and after applying the adjustment

available within the assessment; and hence a higher amount of redundancy that is not accounted for by Relief. On the other hand, MRMRD combined with FFS and linear regression has the second best accuracy, compared to AFSP, across most of the risk values.

Since AFSP offers a trade off between redundancy and relevance, in contrast to DFSP, the quality of the selected feature set is higher and this is reflected in an increase in prediction accuracy. Linear regression is used in both approaches, but classification is performed differently. While DFSP does not offer a solution for classification, and thus implicitly implements equidistant boundaries, UVSD boundaries used in AFSP reduce prediction error compared to DFSP.

Statistical significance is measured by how much the results of a test deviate from chance. For a multi-class classification problem, a random decision between K classes (based solely on chance) would result in an accuracy around $1/K$ for all classes. In our case for 11 classes, this would result in a percentage accuracy of

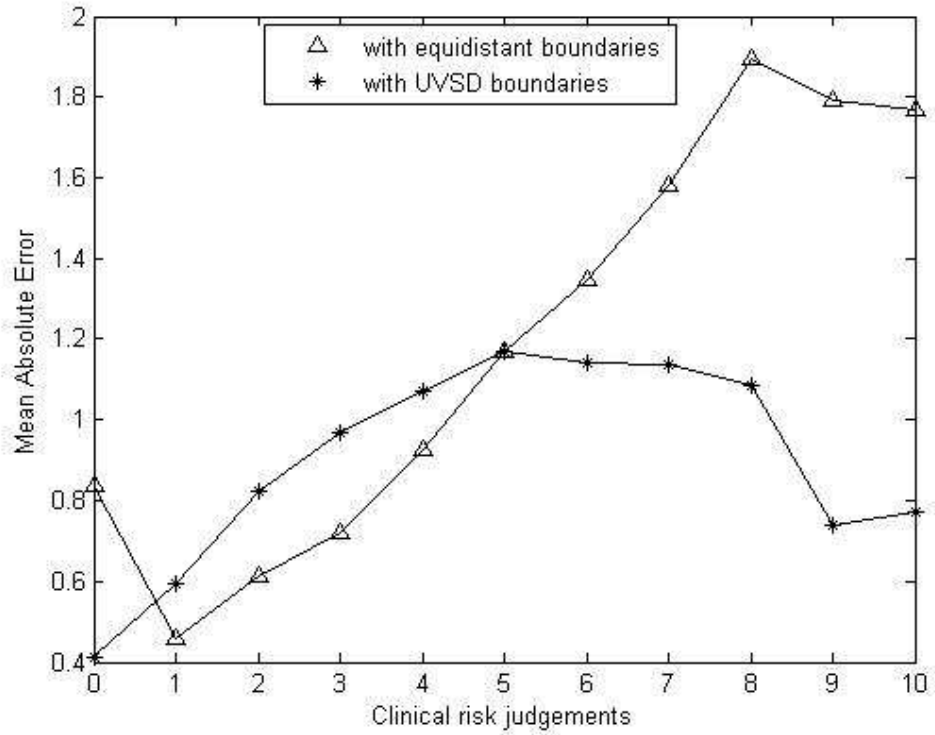

Figure 5.16: The mean absolute error in risk prediction against the clinical risk judgements before and after applying the adjustment

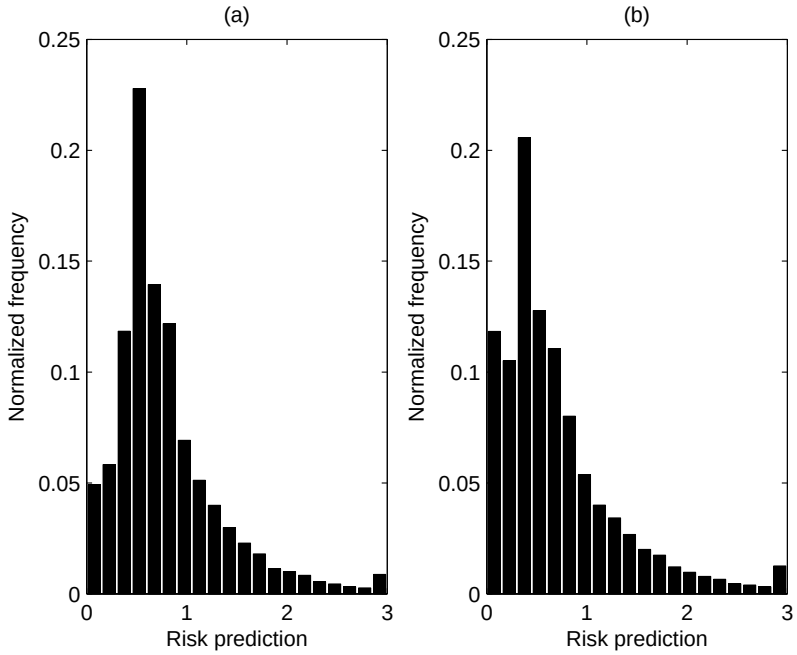

Figure 5.17: Distribution of risk predictions for clinical risk judgement of 0: (a) before adjustment, (b) after adjustment

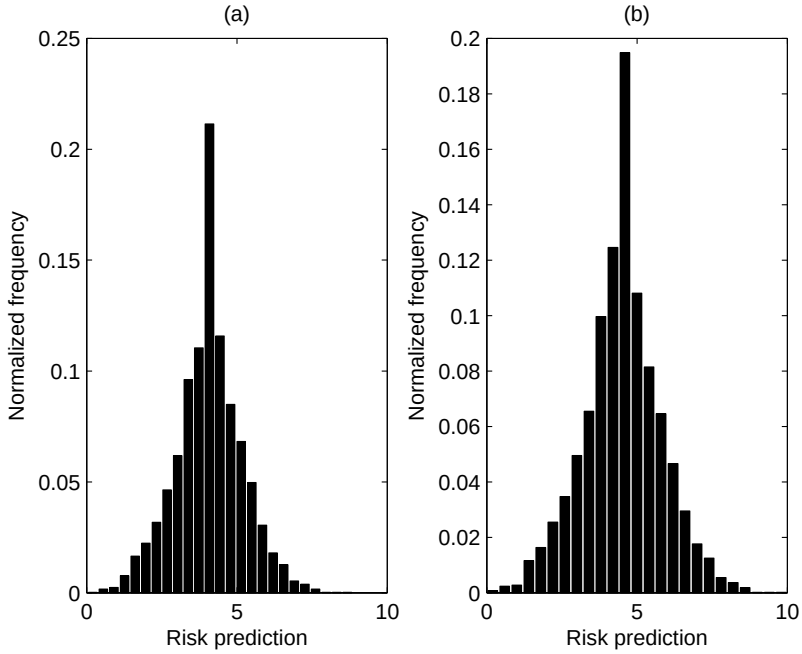

Figure 5.18: Distribution of risk predictions for clinical risk judgement of 5: (a) before adjustment, (b) after adjustment

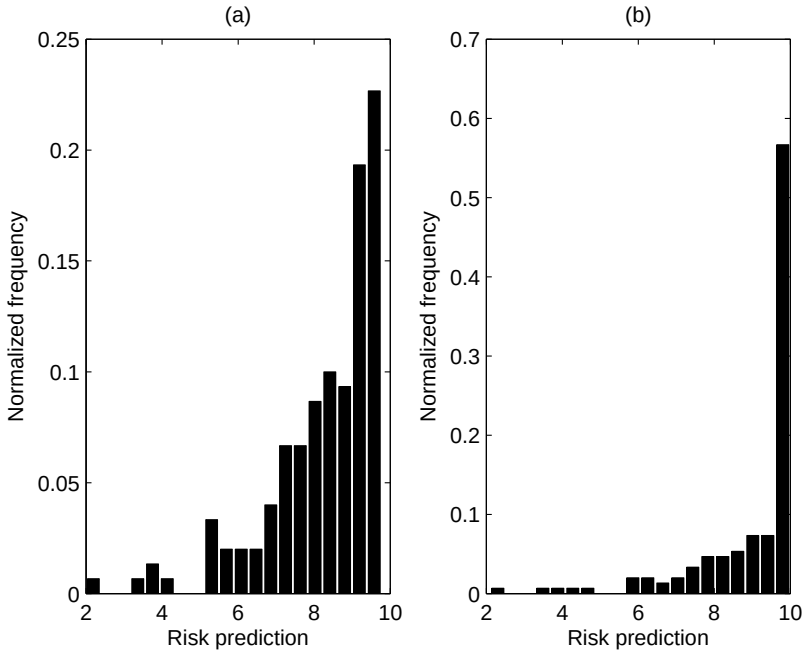

Figure 5.19: Distribution of risk predictions for clinical risk judgement of 10: (a) before adjustment, (b) after adjustment

around 9.09%. For any class an accuracy around that value could be attributed to chance and is not an improvement in prediction accuracy. From Table 5.5, it is clear

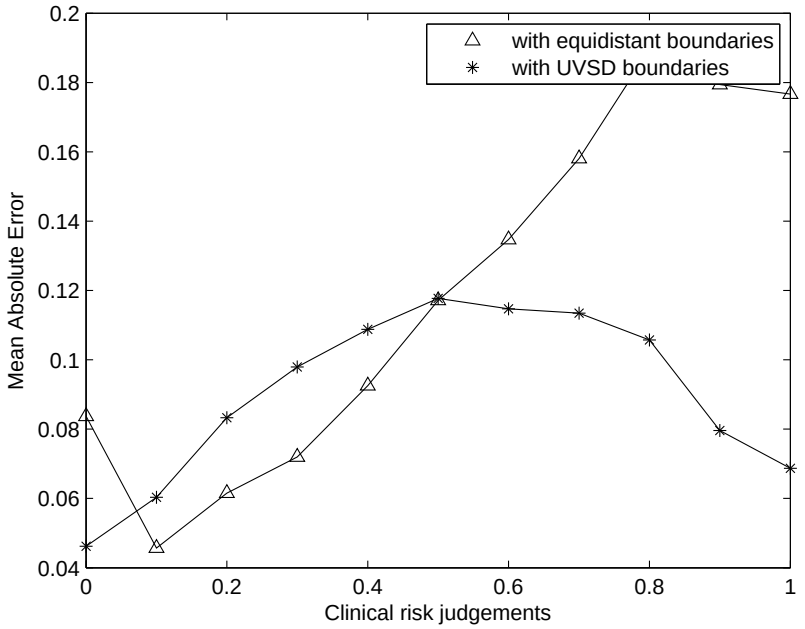

Figure 5.20: The absolute error in risk prediction against the clinical risk judgements with equidistant boundaries and UVSD boundaries

that the results deviate significantly form chance throughout the risk scale. The results show a significant statistical improvement to DTs, RFs and ReliefF, since these have accuracies that are close to 9.09% in the mid range and high risk categories.

5.10 Speed

The AFSP outperforms DFSP with regard to prediction accuracy. However, for implementing AFSP within GRiST, the speed of feature selection and risk prediction needs to be analysed. GRiST is a real-time support system, which means risk assessment speed is important. This has been one of the driving factors in developing the computational optimisations in the algorithm for those elements that need to be executed in real-time while practitioners are conducting assessments. In this section, the speed of feature selection and risk prediction is analysed as different enhancements are proposed.

5.10.1 Search Space Reduction

The preprocessing stage involves applying the correlation threshold, which reduces the size of the candidate feature set and, hence, speeds up feature selection. The concept-descendant exclusion criteria also results in space reduction but not as significant as applying the correlation threshold. Table 5.6 shows that the average time taken for feature selection is significantly reduced when the correlation threshold is applied. The time taken to apply the concept exclusion criteria is insignificant compared to feature selection time, as it represents only 0.2% of total feature selection time, when the correlation threshold is not applied; and 0.4% of total feature selection time when the correlation threshold is applied. Applying the correlation threshold reduces the total time taken for feature selection by almost 3 times and reduces the time taken to filter by ontology by 2.

Table 5.6: Average total time taken and number of clock cycles for feature selection and concept exclusion for a single record with and without applying the correlation threshold

Computation parameters	without correlation threshold		with correlation threshold	
	Concept exclusion	Feature selection	Concept exclusion	Feature selection
Time(s)	0.0010	0.4911	0.0005	0.1753
No. of Clock Cycles ($\times 10^6$)	5.400	2.652	2.700	946.6

5.10.2 MRMR Optimisation

The reduction in the number of summations required at each round of feature selection, discussed in section 3.2.5, results in further reduction of the time taken for feature selection, as shown in Table 5.7. The time taken for feature selection is 1.5 times lower when the summation is optimised and the previous scores are stored, compared to when the total score is computed from scratch at each round.

Table 5.7: Average total time taken for feature selection and number of clock cycles for a single record before and after MRMR optimisation

Computation parameters	before optimisation	after optimisation
Time(s)	0.1753	0.1187
No. of Clock Cycles ($\times 10^6$)	946.6	641.0

5.10.3 Prediction and Classification Time

During feature selection, data with the complete feature set has to be fetched at each iteration, which means that a complete training set will be compiled by the end of feature selection. The training data available from feature selection may be used to calculate regression weights directly, without having to recall the data once more. Table 5.8 shows a slight decrease in weight calculation and prediction time when the data acquisition is not handled during prediction.

Table 5.8: Average time taken and number of clock cycles for computing weights and risk prediction for a single record with and without data acquisition time

Computation parameters	with acquisition	without acquisition
Time(s)	0.0298	0.0218
No. of Clock Cycles ($\times 10^6$)	160.9	117.7

On the other hand, classification and adjustment times are negligible compared to the time taken to compute linear regression weights. Adjustment and classification take on average $3.4X10^{-7}s$ per record, constituting only 0.0016% of the average total prediction and classification time of $0.02180034s$.

5.11 Case Studies

In order to assess the interpret-ability of the results, an analysis of the selected features and examples of assessments and the corresponding predictions are given in this Section. First, we need to verify that the selected features are logically, not only statistically plausible. The 30 most selected cues are listed in Table 5.9.

Table 5.9: A list of the 30 most selected cues for suicide risk prediction with the type of padlock to indicate the frequency at which they change, where "None" means no padlock and that the feature changes frequently, "Silver" padlock is for contextual factors that change occasionally and "Gold" padlock is for historic factors that do not change

Cue	Padlock
Content of suicidal ideation indicates high risk	None
Subject thinks life is not worth living	None
Potential triggers for prospective suicide	None
Current intention to commit suicide	None
Lack of plans for the future	None
Suicidal ideation	None
Lack of regret about trying to commit suicide	Silver
Subject is sad/downbeat	None
Past and current suicide attempts	Gold
Negative feelings about the self	None
Self-harming cuts	None
Distress	None
Time lapse since most recent suicide attempt	Silver
Stage of depression	Silver
Subject's perspective of self worth	Silver
Subject's behavioural presentation during assessment	None
How much did the person want to succeed in suicide attempts	Silver
Potential lethality of suicide method	None
Detrimental changes to relationships	None
Number of suicide attempts	Silver
General personality	Silver
General current behaviour	Gold
Mood swings	None
General social context	None
Chance of discovery after suicide attempts	Silver
Suicide note written for one or more previous attempts	Gold
Motivation and engagement with world	None
Detrimental effects of alcohol misuse	Silver
Adverse life events	Gold
Environment person grew up in	Gold

The selected features, in 5.9 are strong indicators and symptoms of suicide risk [30, 18, 31, 11]. The average correlation to suicide risk of the 30 most selected features is 0.34 compared to an average correlation of 0.18 of all candidate features, which confirms the statistical dependence of the risk on the selected feature set. The features are symptomatic of acute risk episodes which means that they change values frequently from one assessment to the next. The contextual (more persistent) variables that can change values, but not very often, come along occasionally and the historic factors the least. This fits with the idea of identifying and managing the immediate risk issues (the dynamic symptoms) before reviewing the contextual and historic factors that are more to do with risk management. The risk formulation in GRiST mirrors these categories with padlocks, where no padlock means that the features change frequently, silver padlock is for contextual factors that change occasionally and gold padlock is for historic factors that do not change. The type of padlock is shown in Table 5.9, where gold padlocks appear rarely compared to no padlocks.

Two examples are investigated below: one where the prediction is a reflection of the risk factors and the clinical risk judgement (Good Example). and another where the prediction does not reflect the risk profile of the subject (Bad Example).

5.11.1 Good Example

An examples is chosen to display the agreement between risk factors and clinical risk judgement, on one side, and risk prediction, on the other. The example is for a high risk patient, since these are the most critical cases. Table 5.10 shows a strong correlation between: the values of the selected features and their corresponding weights, and the risk prediction. The selected features are all good indicators of suicide risk, with “content of suicidal ideation indicates high risk” and “potential triggers of suicide” having the highest impact on risk. The advantage of the adaptive approach is that the weights and the chosen variables will be particular to each case.

Table 5.10: The selected features, corresponding regression weights (W) and their assessment values (X) along with the risk prediction for a patient with a clinical risk judgement of 10 and a high risk prediction

Selected feature	W	X
Intercept	-0.08	1
content of suicidal ideation indicates high risk	0.40	1
environment person grew up in	0	1
most recent suicide attempt	0.12	1
motivation and engagement with world	0.02	0
stage of depression	0.01	1
general current behaviour	0.01	0
potential lethality of suicide method	0.09	1
frequency of suicidal ideation	0.07	1
regret about trying to commit suicide	0.04	1
distress	0.04	0.8
suicide note written for one or more previous attempts	-0.014	1
potential triggers of suicide	0.29	0.8
Risk Prediction	Prediction (Y_{old}=0.88)	Adjusted Prediction($Y_{new} = 1.00$)

5.11.2 Bad Example

The case in Table 5.11 is classified as a risk of 7 according to the risk prediction, whereas it has been given a clinical risk judgement of 10. The variation may be attributed to two main shortcomings: selection of poor predictors such as “detrimental effects of alcohol misuse” with a correlation of 0.1677 to suicide risk and a large negative estimate of the intercept. When variables with low correlation to the risk are selected, the weights may not reflect the true impact of the features, since the weights of important factors are underestimated. In this particular case, the value of “detrimental effects of alcohol misuse” is 0 which indicates no risk form that factor. However, the choice of features and weights is not based on their particular values in an assessment.

Table 5.11: The selected features, corresponding regression weights (W) and their assessment values (X) along with the risk prediction for a patient with a clinical risk judgement of 10 and a relatively low risk prediction

Selected feature	W	X
Risk Factor (feature)	Weight (W)	Value (X)
potential triggers for prospective suicide	-0.16	1
detrimental effects of alcohol misuse	0.14	0
person's perspective of self worth	0.03	0
self-harming cuts	0.03	0.9
general current behaviour	0.14	0
How much did the person want to succeed	0.03	0.3
time lapse since most recent suicide attempt	0.09	1
number of previous suicide attempts	0.14	0.75
suicide note written for one or more previous attempts	0.02	1
environment person grew up in	0.05	0.9
capacity to cope with major life stresses	0.09	0.9
potential lethality of suicide method	0.05	0.9
impulsiveness	0.06	0.8
chance of discovery after suicide attempts	0.01	0
Risk Prediction	Prediction (Y_{old}=0.68)	Adjusted Prediction($Y_{new} = 0.59$)

5.12 Summary

In this chapter, several feature selection and prediction techniques are compared to the feature selection and prediction components of AFSP. MRMRQ is found to provide the best performance in terms of accuracy compared to other feature selection metrics. Forward selection and FFS provide the best performance among sequential search techniques, but the added computational complexity of FFS outweighs

its benefits. Since the size of the feature set is much smaller than the size of the candidate set, forward selection will be faster to reach a solution than backward elimination. In addition, the results of applying AFSP to suicide risk prediction on data from GRiST are presented. The preprocessing steps have been found to improve accuracy and speed. The linear transformation and UVSD boundaries implemented to adjust for heteroscedasticity not only improve accuracy, but also increase performance consistency across the risk scale.

AFSP outperforms DTs, RFs, DFSP and variants of ReleifF and MLR in suicide risk prediction across all risk values, in terms of accuracy and shifted accuracy.

The enhancements implemented in AFSP result in a significant improvement in processing time and number of computations. Table 5.12 shows that feature selection is the largest contributor to the average total time taken for assessment and the number of computations compared to prediction and classification time.

Table 5.12: Average time taken and number of clock cycles for each component of AFSP per record

Process	time(s)	no. of clock cycles ($\times 10^6$)
Feature selection	0.1187	640.9
Prediction and Classification	0.0218	117.7
Total assessment time	0.1405	758.6

In the following chapter the general applicability of the algorithm will be tested by applying it to other problems. Chapter 6 discusses the applicability of AFSP to predict concepts within suicide risk, such as current intention of suicide and clinical depression.

Chapter 6

Implementation in Sub-Concepts

6.1 Introduction

Clinical judgements of suicide risk rely on two types of cues: information-based cues and judgement-based cues. Some cues such as age, dates and number of attempts are not prone to interpretation and judgement and are simply recorded by clinicians during assessment, while other cues such as current intention of suicide, feelings and emotions and state of mind are assessed by the clinician from a patient's answers. When data is collected by clinicians in their interviews with patients, how their judgements map to the data is obscure. Many attempts have been made to unify and formalize the factors contributing to suicide risk [29, 30, 18, 31], yet the influence of these factors on risk remains highly subjective.

"Current intention of suicide" and "Stage of Depression" are two of the 30 most selected cues in suicide risk prediction as indicated in Chapter 5. In this chapter, an enhancement to suicide risk assessment is proposed, by moderating the impact of sub-concepts on the risk prediction process. First, several risk factors are incorporated to arrive at a more informed estimation of current intention of suicide. This new measurement will help explain the instances where unfortunate outcomes, do not agree with initially stated intentions and risk predictions. Second, predicting clinical depression episodes is addressed, which may help identify patient's in an

episode of depression when the definitive data about the state of depression is not provided.

6.1.1 Concept Nodes

Suicide risk is a high level concept linked indirectly to low level leaf nodes through intermediate sub-concepts (filter nodes). Although, filter questions are discarded in favour of their children when predicting suicide risk, the concepts represented by some filters may be important for several reasons. First, the hierarchical structure resonates with the way human experts structure their risk assessment and management knowledge. Second, interactions between siblings (children of the same parent node) are not accounted for when the leaf nodes are used directly. Third, the constituents of each sub-concept may vary form case to case based on the available data and its relevance. Finally, predicting the severity of certain sub-concepts is important for managing risks as well as assessing them.

Suicidal intent and clinical depression are among the most important factors contributing to suicide risk [125]. Even though they are intangible variables, there are measures that may be used as targets for predicting them [126, 127]. In addition, implementing AFSP for solving a different type of problem, validates how well the algorithm can be generalised.

6.2 Current Intention

Intention is a mental state that represents a commitment to carrying out an action or actions in the future [128, 129]. It is the most influential factor for predicting suicide risk [15], yet its measurement relies on accurate self reports or the clinical judgement of the assessor. The biggest issue facing the accurate assessment of intention is when people deny having intention, because it depends on lack of evidence. While it is reasonable to claim that the presence of current intention of suicide is

an indication of risk, the absence of intention does not necessarily dictate the opposite, since patients may not be open about their intentions [130]. When intentions are misjudged by clinicians, the devised plans and outcomes are not an appropriate match. Therefore predicting intention from other cues in an assessment may be a better measure of intention than a simple Yes/No answer [131].

To illustrate the effect of current intention on clinical risk judgement, the distribution of clinical risk judgements is computed for positive and negative current intention answers. Figure 6.1 shows that the average clinical risk judgement for patients who answered "Yes" is 0.4699 and for patients who answered "No" is 0.1660. This is a highly significant clinical difference, with people being treated very differently with a difference of 3 points on the risk scale. Since 45% of the population with repeat assessments who have answered "No", attempt suicide later on, pointing out unreliability of current intention measure may have prevented the repeat episodes they have effected.

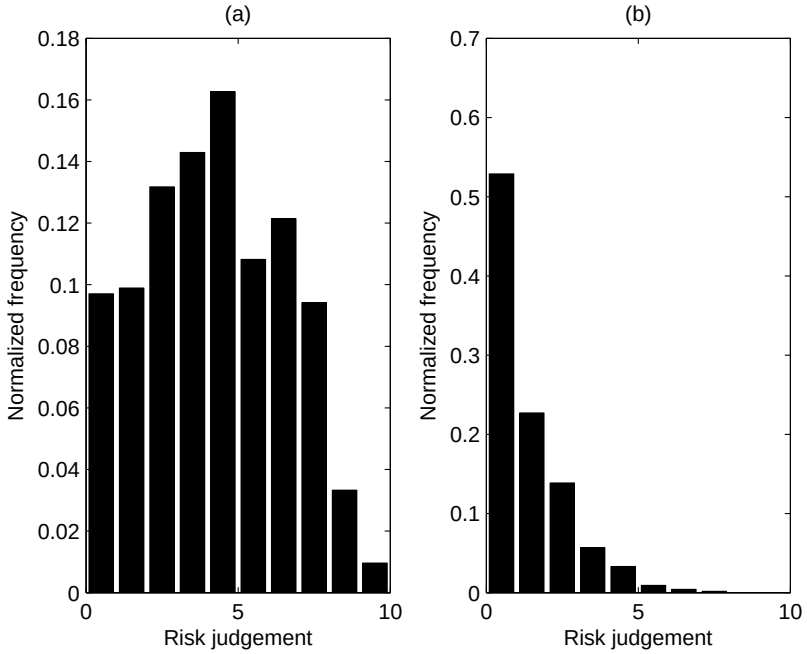

Figure 6.1: Distribution of clinical risk judgement: (a) for patients with "Yes" answer, (b) for patients with "No" answer

In GRiST, current intention is assessed by a filter question (Yes/No answer). The problem arises when the answer is "No", since it is effectively categorical, compared

to a “Yes”, as “Yes” opens up a number of subsequent questions and alerts the assessor to the risk of attempting suicide. The goal is to find out whether it is possible to detect whether the negation of current intention is actually correct or not. In other words, when the current intention answer is “No", its reliability is measurable if differences exist between people in the GRiST database who have answered “No” to having current intention of attempting suicide, but made an attempt later, and those who answered “No” but did not attempt suicide. First, all the patients with “No” answers to current intention and who have a subsequent assessment are compiled. The answer in an assessment is labelled as reliable or unreliable by inspecting its subsequent assessment to find out whether a patient has attempted suicide since the assessment in question. This produces two classes: patients with an unreliable “No” answer who are labelled as class C_1, and those with a reliable “No” answer, class C_2. The total number of assessments is 100450, of which 13429 assessments are eligible for current intention classification. The size of C_1 is 6066 assessments and the size of C_2 is 7363 assessments. The average time lapse between subsequent assessments used to assess intention is 22 days, with a minimum of 0 days (same day assessment and attempt) and a maximum of 100 days between assessment and later attempt. It is true that they may not have had the intention at the assessment time, especially if that assessment was not in the near past. However, the lack of intention was not robust and, in that sense, unreliable as there were issues that may have been picked up on earlier, if intention was assessed from the data.

The following subsections tackle the problem of predicting current intention of suicide using two different approaches; fixed [116] and adaptive feature sets. The first approach is based on a fixed feature set followed by linear regression, while the second approach utilises the preprocessing and feature selection components of AFSP, followed by linear regression. A fixed feature set is tested first, since [116] shows that it offers superior performance in this particular problem. However, the disadvantages of using a fixed set outweigh the benefits, as will be shown through the results. Linear regression is favoured over logistic regression for the sake of speed, as these concepts need to be predicted in real-time and in parallel to risk assessment and prediction. Section 6.2.3 compares the results of both approaches from several aspects.

6.2.1 Fixed Set

The fixed feature set is chosen to maximise the separability between C_1 and C_2 for each feature. The MLE of the mean and variance of each feature in each class is calculated assuming a Gaussian distribution. Then the correlation between the feature and the DV (the classes) is calculated for each feature to produce the relevance measure D_m for feature X_m.

The relevance is limited by the sample size available for each feature as shown in Table 6.1, since selecting more features reduces the sample size. The problem with this approach is that only records that have the complete feature set may be classified.

Table 6.1: The top 20 features (X_m) listed in order of relevance (D_m), the number of occurrences (N_m) of the cue within the sample and percentage of assessments for which the cue is missing

X_m	D_m	N_m	Missing (%)
Most recent suicide attempt	0.5929	9009	32.91
Suicide attempts escalating in frequency	0.2872	6247	53.48
Strength of suicidal ideation	0.1308	3236	75.90
Content of suicidal ideation indicates high risk	0.1295	3294	75.47
Potential triggers of suicide	0.1038	5909	56.00
Life not worth living	0.1027	6662	50.39
How many suicide attempts	0.0987	6790	49.44
Helplessness	0.0979	6588	50.94
Distress	0.0973	6842	49.05
Plans for the future	0.0939	6690	50.18
Anxiety-based emotions	0.0926	6901	48.61
Worthlessness	0.0869	3925	70.77
Sad/downbeat	0.0853	6836	49.09
Sleep disturbance	0.08362	4153	69.07
Angry emotions	0.0806	6878	48.78
Ability to control suicidal ideation	0.0798	3364	79.85
Negative feelings about the self	0.0776	6628	74.95
Perception of the service received	0.0772	1342	50.64
Concordance	0.0762	1391	89.64
Impulsiveness	0.0621	3963	70.49

After the set is chosen, regression weights are computed from training data. Since a fixed set is used, the regression weights will be the same for the entire test set and

are only computed once, which makes the fixed set approach faster. Four experiments are conducted using different fixed sets. First, the most influential variables are selected and sets of 1, 2 and 3 variables are chosen based on relevance. Second, sample size is taken into consideration in experiment four in an attempt to reduce the number of drop-outs. Drop-outs are records that do not have a complete set of the features and thus may not be classified using this approach. Table 6.2 summarises the chosen features in experiments 1-4.

Table 6.2: Top seven cues in order of relevance and the chosen feature sets for experiments (Exp.) 1-4

Cue	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Most recent suicide attempt	✓	✓	✓	✓
Suicide attempts escalating in frequency		✓	✓	
Strength of suicidal ideation			✓	
Content of suicidal ideation indicates high risk				
Potential triggers of suicide				✓
Life not worth living				
How many suicide attempts				✓

6.2.2 AFSP Parameters

The parameters required to apply AFSP to current intention are the correlation threshold and the score threshold. These parameters are computed by conducting two separate golden section searches. The correlation threshold ρ_{th} that minimises MSE is 0.05, which cuts down the candidate set size to a maximum of 49 and an average of 28. A value of $v_{th} = 21$ for the score threshold is found to minimise MSE when applied in conjunction with the previously computed correlation threshold. The average resulting set size, when these conditions along with concept exclusion are applied, is 7.72.

6.2.3 Current Intention Results

Performance is measured through plotting the Receiver Operating Characteristic (ROC) curve and comparing the results for various experiments with fixed and adaptive sets.

6.2.3.1 ROC

The output prediction Y from linear regression is compared to a threshold λ, such that if $Y > \lambda$, then class C_1 is chosen (unreliable "No" answer), otherwise C_2 is chosen (reliable "No" answer). The ROC curve is plotted for both approaches, showing the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold values. The line of no discrimination marks the break-even point, where the results are as good as a random binary choice. The ROC curve significantly deviates from chance in both approaches. A comparison between Figures 6.2, 6.3 and 6.4 shows that the most influential variable offers a superior FPR compared to the other sets, at low values of TPR. However, the area under the curve is lowest, since FPR deteriorates quickly as the TPR increases. Figures 6.3 and 6.4 exhibit a larger area under the curve than Figure 6.2, but have a higher FPR at the maximum accuracy point. Experiment four does not offer a performance advantage, as shown in Figure 6.5, but targets the number of drop-outs in an aim to reduce it. On the other hand, the performance of AFSP, illustrated in Figure 6.6, is worse for low values of FPR, but achieves the best performance for high values of TPR and the ROC exhibits the highest area under the curve, compared to the fixed set results.

6.2.3.2 Comparison

In order to compare the fixed set approach and AFSP, the accuracy, TPR, FPR and the number of drop-outs is computed for both approaches. Table 6.3 indicates the superiority of AFSP, with regard to the number of drop-outs. Using the most influential feature does offer maximum accuracy, but is infeasible for 33% of the

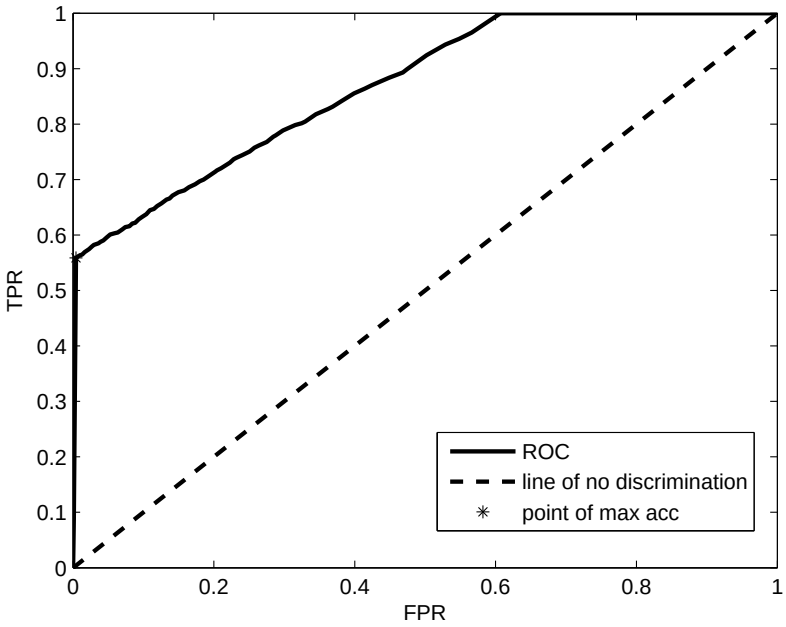

Figure 6.2: ROC curve showing TPR against FPR for different decision thresholds when a fixed set of one variable is used, based on relevance

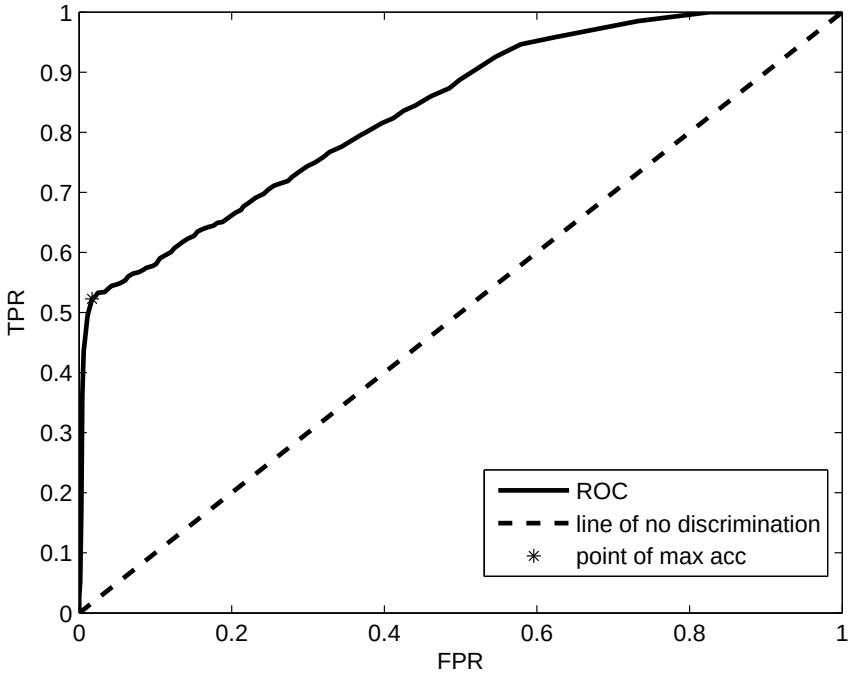

Figure 6.3: ROC curve showing TPR against FPR for different decision thresholds when a fixed set of two variables is used, based on relevance

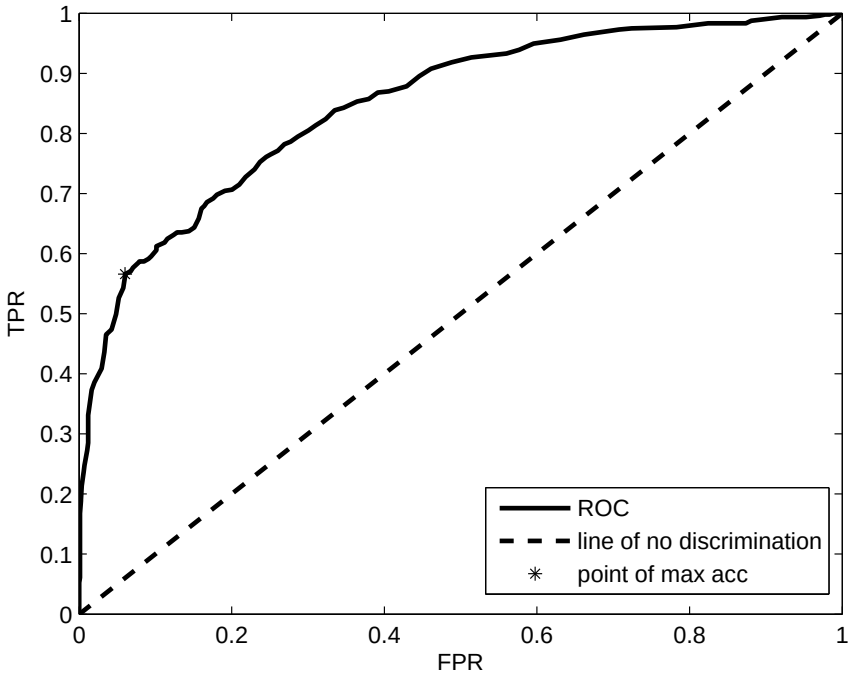

Figure 6.4: ROC curve showing TPR against FPR for different decision thresholds when a fixed set of three variables is used, based on relevance

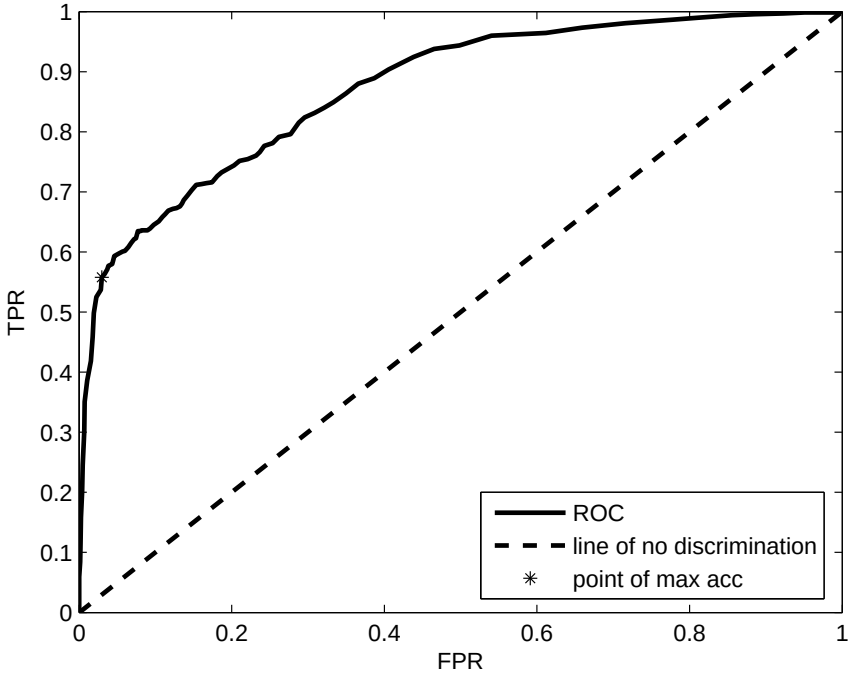

Figure 6.5: ROC curve showing TPR against FPR for different decision thresholds when a fixed set of three variables is used, based on sample size and relevance

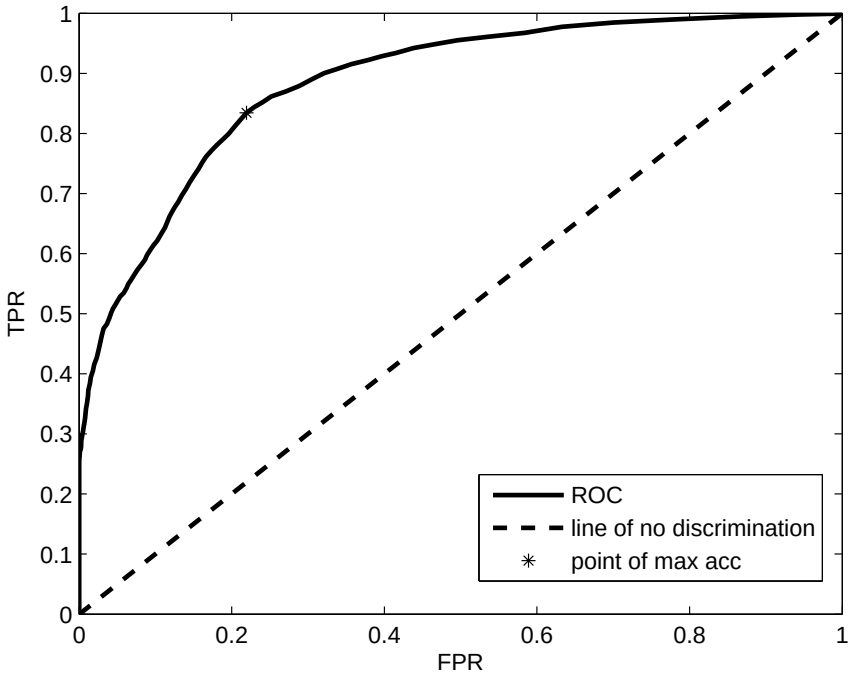

Figure 6.6: ROC curve showing TPR against FPR for different decision thresholds when an adaptive set is used

data. Introducing more features to the fixed set does not improve performance and also increases the number of drop-outs significantly. In experiment 4, when the sample size is taken into consideration along with relevance, the number of drop-outs is reduced and accuracy increases, compared to experiment 3, which uses the same number of features but has a huge number of drop-outs constituting 90% of the data.

Table 6.3: Maximum accuracy (M. acc.) in percentage, TPR and FPR in percentage at maximum accuracy point, and the number of drop-outs (D.o.); for Experiments 1-4 and AFSP

Exp.	M. acc.	TPR	FPR	D.o.
1	87.33	55.88	0.42	4411
2	84.72	52.26	1.71	7856
3	80.53	56.60	6.01	12095
4	86.89	55.77	2.98	10666
AFSP	80.48	83.43	21.94	0

The point with the highest accuracy is not necessarily the best operating point. It depends on a compromise between FPR and TPR. In prediction, TPR is probably

the most important measure, yet FPR is also significant because false alarms may exaggerate the risk and potentially trigger unnecessary interventions. The results in Table 6.3 show the performance of each set when operating at its own maximum accuracy point.

The stability of the performance does not depend on the maximum accuracy point, but is reflected through the rate of change of FPR and TPR when the operating point is varied. To illustrate the balance between TPR and FPR, performance is analysed at almost the same TPR for all set-ups. Table 6.4 shows that AFSP outperforms any fixed set when a performance bound of TPR$> 83\%$ is set, since it offers the highest accuracy and lowest FPR, compared to all other set-ups. This means that when the threshold is varied to favour TPR, the performance of a fixed set deteriorates quickly as it moves away from its optimum operating point. On the other hand, when the threshold is varied to favour FPR, AFSP achieves a TPR of 54.21% at an FPR of 6.21% and an overall accuracy of 79.01%, which is comparable to the performance of fixed sets.

Table 6.4: Accuracy (Acc.), TPR and FPR in percentage when TPR$> 83\%$, for Experiments 1-4 and AFSP

Exp.	Acc.	TPR	FPR
1	68.73	83.13	36.88
2	65.15	83.60	42.56
3	72.75	83.86	33.49
4	72.59	83.14	30.85
AFSP	80.48	83.43	21.94

Logistic regression is also tested using the adaptive feature set and prediction results are compared to linear regression. Besides accuracy, TPR and FPR, Table 6.5 shows that prediction time for linear regression is one eighth the prediction time when logistic regression is used. The reduced processing time does not come at the cost of reduced performance, and thus linear regression should be preferred.

Table 6.5: Percentage Accuracy (Acc.), TPR, FPR and average prediction time (P.t.) in seconds per record at maximum accuracy point when linear and logistic regression are used

	Linear regression	Logistic Regression
Acc.	80.48	79.92
TPR	83.43	80.64
FPR	21.94	21.52
P.t.(s)	0.0156	0.1237

6.2.3.3 Chi-Square Test

In order to verify the significance of the results a Chi-square test is performed to check if the reliability decision is dependent on whether the patient has a repeat attempt or not. The results summarised in Table 6.6 give $\chi^2 = 6367.5$ with a resulting cumulative probability very close to 1, and thus the null hypothesis probability is almost 0, which is lower than the significance level $p < 0.001$. Hence, the null hypothesis is rejected and the actual class and the decision are considered dependent.

Table 6.6: Classifier statistics at maximum accuracy point, C_1 predicts an unreliable "No" and C_2 predicts a reliable "No" for current intention

	C_1	C_2	Total
Repeat	5365	701	6066
No Repeat	1464	5899	7363
Total	6829	6600	13429

6.3 Clinical Depression

Clinical depression or Major Depressive Disorder (MDD) [132] is the most severe form of depression and often results in suicidal tendencies [133, 134]. The question is whether a clear DV for clinical depressive episodes may be used to predict degrees of depression in other patients without a clinical diagnosis. This may then help determine whether interventions relating to depression are appropriate. It may

also help improve the accuracy of predicting suicide risk, since depression is an important predictor of suicide [130].

In GRiST, clinical depression is directly addressed by the questions in Figure 6.7. The current depression status of a patient may be classified into two categories; being in an episode of clinical depression, which is indicated by either "first diagnosis" or "relapse", or not being in an episode, which is indicated by "recovery single episode" or "recovery repeat episodes". The answer for this cue is categorised and used as the label for two distinct classes C_1 and C_2. C_1 denotes currently being in a depressive state, while C_2 indicates the opposite. The number of assessments for which the state of depression is indicated is 47,470, of which 23,770 are in an episode of depression and 23,700 are not.

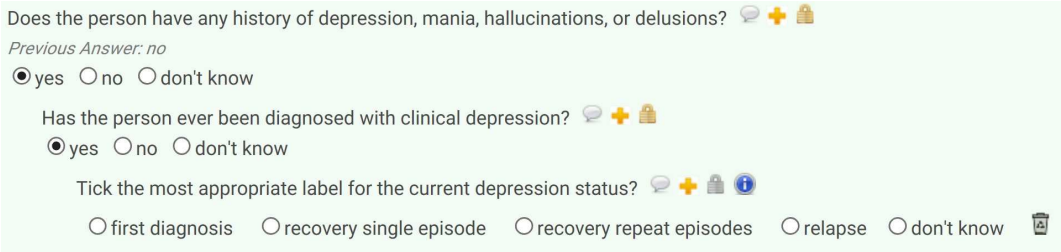

Figure 6.7: Snapshot of depression questions in GRiST

6.3.1 Feature Selection in Depression

The feature selection component in AFSP is applied to the data, as follows. First, correlation between the DV (state of depression) and the IVs (all other cues in an assessment) is calculated. The correlation is used to scale mutual information between variables, to get the individual scores matrix. Correlation threshold and concept exclusion are applied as preprocessing steps and the score threshold is applied as the stopping condition in conjunction with the sample size constraint. The thresholds are determined using two instants of golden section search to minimise MSE. The correlation threshold is found to be 0.1, which reduces the maximum number of candidates to 51 and the average candidate set size to 29. Whereas a score threshold of 14.8 gives the best performance, with regard to MSE. The average size of the selected feature set is 8.45.

6.3.2 Depression Prediction

The selected features are used to compile data from the training set, that has the complete feature set. The training data is used to calculate linear regression weights of the selected features. The data is then classified by comparing the result of linear regression Y to a threshold λ as in 6.1 and 6.2.

$$C = C_1 \; \forall \; Y \geq \lambda \tag{6.1}$$

$$C = C_2 \; \forall \; Y < \lambda \tag{6.2}$$

6.3.3 Depression Results

Table 6.7 shows that the maximum accuracy for predicting depression episodes is 66.91%. To determine the statistical significance of the results, the ROC curve is plotted for various threshold values and a Chi-square test is performed to ensure the dependence of the prediction on the actual class.

Table 6.7: Maximum accuracy (M. acc.), TPR and FPR at maximum accuracy point, in percentage, for predicting depression

M. acc.	TPR	FPR
66.91	65.93	32.11

6.3.3.1 ROC

Figure 6.8 gives the TPR against FPR when the threshold is varied. The threshold at the maximum accuracy point is given by $\lambda = 0.4763$. The ROC curve shows that the decision deviates from chance (represented by line of no discrimination) and, hence, that the classification is statistically significant.

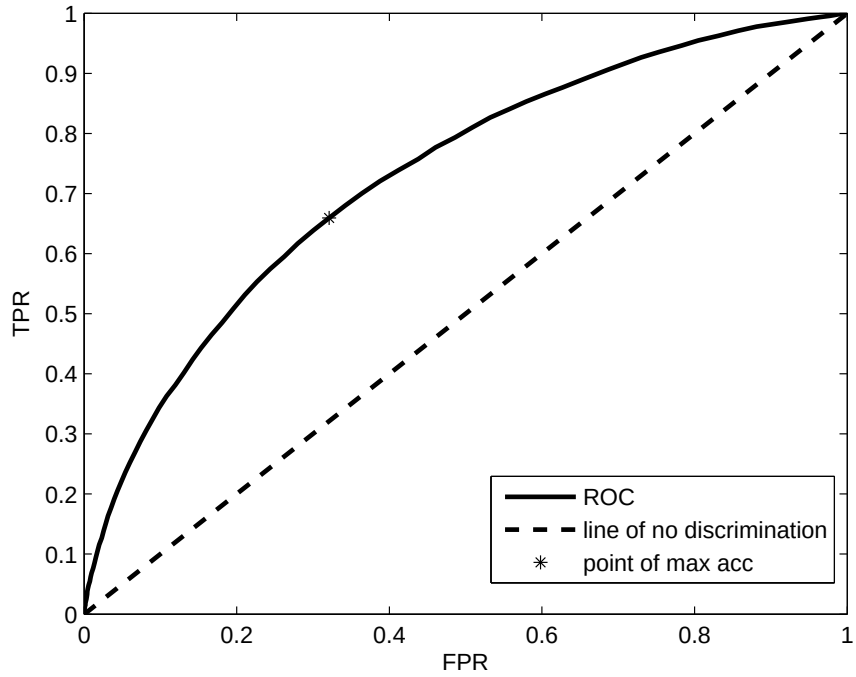

Figure 6.8: ROC curve showing TPR against FPR for clinical depression at different decision thresholds

6.3.3.2 Chi-Square Test

In order to verify the significance of the results a Chi-square test is performed to check if the classification decision is dependent on whether the patient is an episode of depression or not. The results summarised in Table 6.8 give $\chi^2 = 5200.9$ with a resulting cumulative probability very close to 1, and thus the null hypothesis probability is almost 0, which is lower than the significance level $p < 0.001$. Hence, the null hypothesis is rejected and the actual class and the decision are considered dependent.

Table 6.8: Classifier statistics at maximum accuracy point, C_1 predicts a depression episode and C_2 predicts no depression episode

	C_1	C_2	Total
Depressed	15352	8418	23770
Not depressed	7477	16223	23700
Total	22829	24641	47470

6.4 Summary and Conclusion

In this chapter, the feature selection components of AFSP have been applied to two binary classification problems: predicting reliability of current intention of suicide and predicting current clinical depression episodes. An adaptive feature set yields competitive performance when compared to a fixed set in predicting current intention reliability with regards to prediction accuracy. Moreover, an adaptive set accounts for all cases, in contrast to a fixed set, where the number of drop-outs increases with the number of features in the set. Overall, the results of both prediction problems are statistically significant. The significance of the results to suicide risk assessment is summarised below.

Most existing measurements of current intention are effectively categorical, with the outspoken statement denying intention considered reliable. We have used AFSP to measure the reliability of these negative statements, which, in turn, may be used to alter the risk assessment process. Records with unreliable "No" answers, have been missed out by clinicians and DSSs. In such cases, the clinician's judgement of intention does not reflect the patient's true intention, which is manifested through the patient attempting suicide afterwards. Using an adaptive set of cues, selected through applying AFSP, a large portion of patients; who have gone through a suicide attempt while their current intention statement dictated the opposite, has been predicted. Unreliability of current intention could improve risk assessment by raising a flag to alert assessors to collect more data, and thus a well-informed decision is made regarding a patient's case.

On the other hand, a clinical diagnosis of depression may not be disputed. However, in many cases predicted levels of depression in people who are not in an actual clinically-diagnosed depressive state can help improve risk prediction and risk management. The results highlight how depression may be predicted from the available set of cues with high accuracy which may be used when a clinical assessment of depression has not been previously conducted.

Chapter 7

Deployment in Clinical Practice

7.1 Overview

When an assessment is submitted, a part of the back-end process is for GRiST's consensual risk evaluation module to examine the assessment and find similar assessment profiles in the database. The risk levels assigned by the clinician to the top-level risks are then compared to those in the similar assessment pool. If there is a large deviation between the risk judgement and consensus, the user is given an opportunity to amend/explain the values they originally provided [135].

AFSP will be used to perform automated predictions in place of GRiST's consensual risk evaluation module and provide a weighted set of explanatory factors for the prediction. The predictions will not be provided in advance of the clinical judgements and are not intended to replace them.

According to the Evidence Standards Framework for Digital Health Technologies (DHT) [9], tools that perform calculations with an impact on treatment, diagnosis or care are classified as Tier 3b [9]. Consequently, if AFSP is to be deployed in clinical practice through GRiST, to aid clinical decision making, the evidence for effectiveness standard for Tier 3b tools needs to be satisfied. In addition, the tool has to meet the evidence for effectiveness standards for Tiers 1 and 2 [9].

7.2 Evidence for Effectiveness Standards

Since, GRiST data involves vulnerable groups, such as at-risk-adults, the best practice standard has to be met for all Tiers [9]. The following Section details the evidence provided for meeting effectiveness standards for Tiers 1 and 2. Section 7.2.2 explains the design of a Randomized Controlled Trial (RCT) to meet the requirements for the best practice standard for Tier 3b tools.

7.2.1 Tiers 1 and 2

Tables 7.1 and 7.2 describe the best practice standard and the evidence provided for meeting the standard.

Table 7.1: Evidence categories for Tier 1 tools, with a description of the best practice standard [9] and evidence of meeting the standard

Category	Best Practice Standard	Presented Evidence
Credibility with UK health and social care professionals	Published evidence documenting the role of relevant UK health experts in the design and development of the DHT	GRiST [24] structure is based on a model of expert knowledge in the mental health domain, elicited by interviews and has been built to facilitate the generation of automated risk predictions [45]

Relevance to current care pathways in the UK health and social care system	Evidence to show successful implementation of the DHT in the UK health system	GRiST is currently used by the following mental health organizations within the UK [24]: • Worcestershire Health and Care NHS Trust • Birmingham Children's Hospital NHS Foundation Trust • Cumbria Partnership NHS Foundation Trust • Humber NHS Foundation Trust

Acceptability with users	Published evidence to show that users are satisfied with the DHT	Service users and practitioners have played an essential role in the development and improvement of GRiST, examples of work integrating users' feedback into GRiST are: • An evaluation of the clinical implementation and adoption of GRiST [25] • Integration of service users and practitioners expertise within a web-based environment [46] • Developing GRaCE-AGE a component of eGRiST for self-assessments by older adults [136]
Accurate and reliable measurements	Analysis which shows that the data generated by the DHT is accurate and reproducible and that clinically relevant responses are detected	Evidence has been presented for the reliability and repeatability of the clinical risk judgements and the accuracy of risk predictions in Chapter 5
Accurate and reliable transmission of data	Technical data showing that numerical and text information is not changed during the transmission process	Assessment data is collected by collaborators in mental health institutes and is transmitted without modification to GRiST servers for storage

Table 7.2: Evidence categories for Tier 2 tools, with a description of the best practice standard [9] and evidence of meeting the standard

Category	Best Practice Standard	Presented Evidence
Reliable information Content	Evidence of endorsement, accreditation or recommendation by NICE, NHS England, a relevant professional body or recognised UK patient organisation	Peer-reviewed funding from health organisations, including: • The Health Foundation • NIHR Many health organisations have been involved in developing GRiST or have paid licences to use it, including: • Humber NHS Foundation Trust • City Healthcare Partnership, Hull • Cumbria Partnership NHS Foundation Trust • Birmingham Children's Hospital • Worcester NHS Trust • Orkney NHS Trust • Barchester Healthcare • Raphael Healthcare • Rossie Young People's Trust • Northern Healthcare

Ongoing data collection to show usage and value of the DHT	Evidence that data on usage and user satisfaction is being collected in line with the minimum standard and can be made available to relevant decision-makers	Data on user satisfaction and feedback from clinicians is continuously collected through workshops and training sessions for professionals using GRiST in practice [24]
Quality and safeguarding	Show that appropriate safeguarding measures are in place around other communication functions within the platform and describe who has access to the platform and their roles within the platform	Personal identification information is held on a separate server to the mental-health data with access from patient records to the mental-health data using a one-way salted hash function. Moreover, the data is handled within GRiST in accordance to the General Data Protection Regulations (GDPR).

7.2.2 Randomised Controlled Trial

There is recognition in the health informatics research community that health technology is not easily evaluated using RCTs [137]. However, RCTs do provide the most convincing evidence, if it can be collected and so we will devise an RCT that would be possible for evaluating the use of AFSP within GRiST.

Since GRiST, in isolation, does not perform calculations of risk, only AFSP is considered as a Tier 3b [9] tool. Hence, the target of the RCT is not to demonstrate the usefulness of GRiST in the general sense, but to test the impact of AFSP on clinical decision making and corresponding outcomes when implemented within GRiST. The design of such RCT is detailed hereunder.

An important element of designing RCTs is to define a clinically relevant objective a priori [10]. In our case, the end-goal is to establish whether or not risk predictions and corresponding weighted risk factors, generated by AFSP, improve patient outcomes with regard to suicide risk. Table 7.3 summarizes the proposed design of an RCT to evaluate the effectiveness of AFSP and the flow of the RCT is displayed in Figure 7.1.

Table 7.3: Summary of RCT design elements [10] for evaluating the effectiveness of AFSP within GRiST

Hypothesis	Risk Predictions and weighted explanatory factors generated by AFSP improve patient outcomes with regard to suicide risk
Study Design	Parallel [10]; i.e. the sample will be divided into two groups: one for which AFSP is used by clinicians (Group **A**), and another for which AFSP will not be used as part of the assessment (Group **B**)
Allocation Ratio	1:1; i.e. the two test groups will have the same number of samples
Procedure	• Group **A**: For each subject upon the completion of a new clinical assessment by a clinician using GRiST, an automatically generated value of the risk prediction will be generated by AFSP along with the corresponding feature set and their weights. The clinician is intended to use this information to aid in decisions regarding risk judgement and management plan. • Group **B**: For each subject a clinical assessment will be conducted using GRiST and the corresponding clinical risk judgement and management plan will be given by the clinician without the use of any decision support. For both groups when subsequent assessments are conducted, the outcomes will be measured.

Outcomes	• Number of suicide attempts following the initial assessment • Direction of travel of risk indicated by Clinical risk judgements in subsequent assessments
Randomisation Method	Covariate adaptive randomization[138], where a new patient is assigned to **A** or **B** by considering previous assignments of subjects with the following covariates: • Age group (children and adolescents, adults and older adults) • Gender All subsequent assessments of the same patient need to be in the same group, following the first assignment of the patient.
Allocation	When an assessment is being conducted covariates will be collected during the assessment and allocation will be performed on line depending on covariates of previous members of both groups and their relative size

Blinding	• Patients will not be aware of group assignments. However, patients will be informed of the trial and that they may be in one of the two groups and their consent will be obtained prior to participation. • Clinicians will be made aware after the assessment is completed and a clinical risk judgement is provided for the patient, to avoid any bias in the collected data upon which the risk prediction is based. • Clinicians will not know whether AFSP is turned on or off, unless it is turned on and the AFSP detects a deviation between the judgement and the prediction.
Statistical Analysis	Chi-square test of significance is used to assess whether changes in outcomes are dependent on patient groups

7.3 Monitoring and Maintenance

The system will be monitored and maintained to ensure that assessments run smoothly and accurately. Monitoring of the speed of conducting and completing assessments and risk predictions is necessary to ensure no delays occur due to high demand on processing time. Potential upgrades to the processing capabilities and memory should be investigated to maintain Quality of Service (QoS) parameters below predetermined thresholds.

Feedback from service users is an important aspect of updating and improving the system to meet users' expectations and needs. In addition, autonomous and continuous parameter update is embedded within AFSP to ensure the integrity and reliability of global parameters.

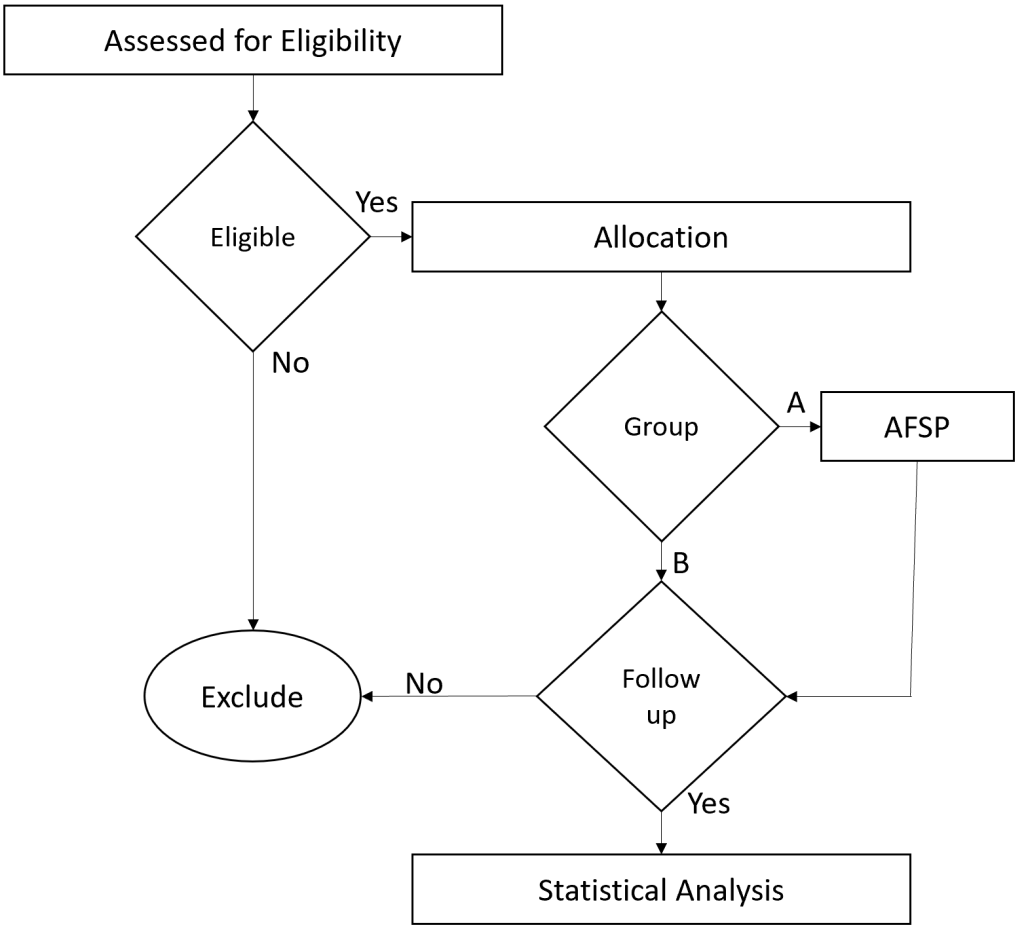

Figure 7.1: Flow chart of the proposed RCT

7.3.1 Quality of Service

A number of QoS parameters may be used to measure system performance and ensure a good user experience. Potential QoS parameters include processing time and number of crash reports. The description of QoS parameters in Table 7.4 details their definitions and how they may be measured.

Table 7.4: QoS parameters and their description

Processing time	time taken to calculate risk predictions and output the results to clinicians
Number of crash reports	number of times the system fails to retrieve, store or process data

7.3.2 Feedback

Feedback from users is crucial to the development of GRiST and all corresponding components including AFSP. The two main sources of feedback are assessors, including clinicians and para-professionals, and patients being assessed by clinicians using GRiST with AFSP enabled. Clinician feedback may be captured through questionnaires and interviews with clinicians using the service. Feedback from patients is difficult to collect and attribute to the use of AFSP or GRiST, but may be measured indirectly through changes in patient outcomes or reports from clinicians on patient experience.

7.3.3 Global Parameters Update

As more data is collected through GRiST, autonomous update of global parameters will be periodically performed to ensure the values reflect the newly added data. Table 7.5 lists global parameters and the corresponding update processes.

Table 7.5: Global parameters and description of update processes

Category	Parameters	Process
MRMRQ parameters	Relevance D	• Calculate correlation between risk judgements and cues
	Redundancy R	• Calculate probability distributions and joint probability distributions of cues • Calculate mutual information between cues
	Individual Scores S	• Divide rows in Redundancy R by corresponding cue Relevance D
Preprocessing Parameters	Correlation Threshold	• GSS to compute correlation threshold
Adjustment and Classification	Adjustment Parameters α and β	• Calculate risk predictions for all available data • Run auxiliary regression between predictions and clinical judgements
	Decision Boundaries λ	• Calculate MLE estimates of μ and σ^2 • Calculate λ based on UVSD

Chapter 8

Conclusion and Future Work

In this thesis, the Adaptive Feature Selection and Prediction (AFSP) algorithm was introduced, to address mental health risk assessment from survey based data. The algorithm was devised to tackle four different issues, that are common in the mental health domain; handling missing data, reducing dimensionality and complexity, autonomously updating parameters and eliminating heteroscedasticity.

An introduction to the topic of mental health risk assessment, including clinical assessment and decision support systems, is given in Chapter 1. The GRiST DSS was also introduced and the basic idea of AFSP was presented.

Chapter 2 reviewed approaches to mental health risk assessment, including expert-based, computer-based and hybrid systems. The structure of GRiST and data-related issues were thoroughly investigated. A survey of feature selection, prediction and classification techniques was included in this chapter. The idea of the DFSP algorithm, the predecessor to AFSP, was explained and possible areas of improvement were highlighted.

In Chapter 3, AFSP was introduced, beginning with the rationale behind it. The development of the feature selection algorithm was described along with the prediction and classification method. A linear adjustment and boundaries based on unequal variance were suggested to overcome heteroscedasticity. A method for autonomous parameter update was also introduced in this chapter.

The implementation of AFSP, within GRiST, for predicting suicide risk was explained in Chapter 4. The parameters needed for feature selection, prediction and classification of suicide risk were detailed with a distinction between off-line and on-line parameters.

The results of suicide risk prediction using AFSP were presented in Chapter 5. Prediction accuracy, shifted accuracy and processing time were used to measure performance. The optimisations introduced through preprocessing and in feature selection reduced assessment time and enhanced accuracy. Finally, a comparison between AFSP and several other methods, with regard to prediction accuracy, emphasised the superior performance of AFSP.

In Chapter 6, AFSP was applied to sub-concepts within suicide risk, namely current intention of suicide and clinical depression. The aim of predicting current intention reliability, was to determine whether intention may be better assessed through risk-related cues, rather than being enquired about directly or evaluated by the assessor. The results showed that AFSP outperforms a fixed set approach, with regard to the number of drop-outs and the area under ROC curve. On the other hand, predicting clinical depression was more challenging yet achievable. The statistical significance of the prediction of depression was established through plotting the ROC curve and performing a Chi-square test of significance.

Chapter 7 proposes a design for an RCT to facilitate the deployment of AFSP in clinical practice. In addition, means of monitoring and maintenance are suggested and QoS parameters are investigated.

Overall, AFSP handled the shortcomings of the data and the domain. Each suggested component improved one or several aspects of performance. The algorithm generalised well when used to tackle diverse problems. The following section explores the issues and how well they have been addressed in more detail for the components of AFSP, which leads to areas requiring further research.

8.1 AFSP Summary and Discussion

The AFSP was divided into three main stages; preprocessing, feature selection, and prediction and classification. The preprocessing stage involved applying a threshold on correlation and concept exclusion. The correlation threshold was applied to suppress features with low correlation to the DV, to cut down the number of candidate features and improve the quality of candidates. While the threshold did not significantly affect accuracy, it resulted in a reduction in feature selection time to one third of the original time. The correlation threshold was computed off-line using a golden section search.

The concept-descendant exclusion criteria was used to discard filters when their children were present, which significantly improved prediction accuracy for high risk categories. Concept exclusion was particularly effective for high risk values, since high risk patients were usually investigated rigorously, and thus had more data and more children present in their records.

MRMR was chosen for feature selection to provide a balance between relevance to the DV and redundancy within the set. Linear correlation with the DV was used as the relevance parameter and mutual information between IVs was used as the redundancy measure. The MRMR scores, used for selection, were based on the quotient of the redundancy and relevance measures, rather than on difference, to overcome normalisation issues and avoid introducing new parameters for normalisation. Instead of using exhaustive combinatorics, a feed forward approach was implemented to facilitate running the algorithm in real time. The stopping condition for feed forward selection was based on the combined score of the set, rather than on the traditional MSE, which transformed the approach from a wrapper method, in which the predictions had to be performed over the training set at each iteration, to a filter method, in which selection depends on predetermined parameters. The stopping condition was determined off-line by performing a golden section search to minimise MSE in the predictions, compared to clinical risk judgements. A constraint over the sample size available for training, was applied in conjunction with

feature selection, to ensure the available data for a feature set is sufficient to train a prediction model reliably.

Linear regression was preferred over other prediction methods for two main reasons. First, linear regression models provided explanations that may be easily interpreted by clinicians, since the weights can be directly input to the GRiST cognitive model to provide an explanation that resonates with the way experts structure risk assessments. Second, calculating regression weights was straightforward and did not require any iterative training, which was advantageous to real time implementation. Since the average time taken for computing the weights and the risk prediction per record was less than one sixth of the average time taken by AFSP to assess one record, regression weights were computed on-line for each assessment and were not stored for later use.

The nature of mental health data led to a violation of a number of assumptions incorporated in linear regression analysis. First, the boundedness of the DV and the non-uniform distribution of training data over risk categories, led to heteroscedasticity. The presence of heteroscedasticity was proved by performing a Bruesch-Pagan test. The parameters for the linear transformation of the predictions, applied to reduce heteroscedasticity, were autonomously computed by performing an auxiliary regression. The adjustment boosted performance at both ends of the risk scale.

Since clinical risk judgements were discrete, risk prediction was, in essence, a classification problem. Computing continuous predictions rather than directly performing classification, reduced the dimensionality of the classifier, since classification was dependent on the prediction and not the features. Performing classification over linear regression predictions, transformed the problem from a multinomial multi-class problem to a number of single variable binary classification problems. To account for heteroscedasticity, decision boundaries were chosen to minimise probability of error in an unequal variance setup. A comparison between the suggested UVSD boundaries and equi-distant boundaries showed a significant increase in accuracy when UVSD boundaries were used. While the boundaries were calculated off-line, using logistic regression for classification would have required train-

ing a classifier on-line, which would have added a significant amount of processing time.

When compared to its predecessor, Nagy's DFSP [1], AFSP performed better with regard to prediction accuracy across all suicide risk categories. This is mainly attributed to the enhancement in the feature selection component of AFSP, that provided a balance between redundancy and relevance, in contrast to DFSP which emphasised relevance over redundancy. Moreover, an autonomous method for learning AFSP parameters has been devised, while DFSP parameters were manually chosen. Finally, the classification component of AFSP enhanced the categorisation of predictions into risk categories, whereas no means for classification were suggested in [1] for DFSP predictions. AFSP was also compared to other methods and when several variants were used for each component. The proposed method outperformed all other methods and had more consistent and stable results across the risk scale.

Even though the algorithm was tested using only data from one database, GRiST [24], the data-related issues addressed by AFSP are common to the entire domain. In addition, the parameters required to run AFSP are fully computable and have not been set empirically , such that the method may be easily implemented for different data, mental health risks and/or databases.

8.2 Limitations and Future Work

Although AFSP outperformed other approaches on several levels and in different areas, it still had a few limitations. The following subsections explain the limitations and propose future improvements to tackle them.

8.2.1 Parameter Learning

Although the parameters in AFSP were autonomously computed, they were learnt in isolation. While each parameter was computed, other parameters were kept constant or discarded. This one-at-a-time approach was used to simplify parameter learning. In addition, since parameters were computed using different methods, including; golden section search for the thresholds, linear regression for the adjustment and UVSD for the boundaries, the effect of one parameter over the other was impossible to establish.

Further investigation of parameter estimation techniques is needed, to find a method of optimising the entire set of parameters in parallel. However, due to the high dimensionality and intricate nature of the data, sophisticated parameter estimation techniques will require a huge amount of time to converge. Hence, the optimisation of parameter search techniques will be crucial to the feasibility of simultaneous parameter estimation.

8.2.2 Memory Requirements

Since GRiST has an ever-growing database of clinical assessments, the amount of available training data increases by the hour. During off-line parameter estimation and on-line assessment, linear regression weights are repeatedly computed from training data. Although, the time taken to compute weights and prediction is insignificant, the memory requirements are not. On-line, multiple assessments may run concurrently, while off-line, multiple regression models are computed during parameter estimation. The amount of training data compiled in memory may become infeasible, as multiple processes are running.

So far, AFSP introduced a lower bound on the amount of data available for training, to ensure weights were accurately computed. An upper bound on the sample size, that is a function of the number of features may be applied in conjunction to ensure that memory is not exhausted. Another approach that may be used is statistical al-

location of memory depending on the requirements of each running process, which would provide a somewhat adaptive upper bound on the size of training data, that will depend on the amount of activity on GRiST at a time.

8.2.3 Clinical Judgements

Clinical risk judgements were used as targets or labels for developing and training AFSP. The predictions provided do not represent a specific clinician, since parameters were computed over the entire set, and thus represent consensus. However, the accuracy of each prediction is tested against the clinical risk judgement provided for a particular assessment. The judgement for each assessment is provided by only one clinician and may easily deviate from consensus. Moreover, judgements may have systematic bias that would emerge in consensus as well, and thus are not a perfect representation of risk.

A better measure of the accuracy of risk prediction would be through tangible outcomes. Mental health risks like suicide, are low base rate events and thus measurable outcomes are not widely available. Intermediate outcomes, though, like repeat suicide attempts or monitoring the increase or decrease of certain symptoms may provide better validation for risk prediction. However, monitoring intermediate outcomes require repeated assessments of the same patient and the quality of the outcomes and feedback through the system will improve with the number of repeat assessments for a patient.

8.2.4 Detecting Outliers

Outlier detection algorithms are used in conjunction with linear regression to improve the robustness of the model [139]. The presence of outliers in training data, used to compute regression weights, may affect the performance of the model and lead to over-fitting. However, applying an outlier detection algorithm is not only time consuming, but may also eliminate a large number of records, since mental

health data is heterogeneous. In addition, a new parameter is needed for outlier detection, which would increase the number of parameters required to run AFSP. Incorporating an outlier detection component within AFSP needs to be tested to weigh its merits with regard to performance, if any, against added complexity and parameters.

8.3 Final Conclusion

In this thesis the Adaptive Feature Selection and Prediction (AFSP) algorithm has been proposed to enable real time risk prediction from mental health assessments conducted by clinicians. The components and parameters of the algorithm have been discussed and the performance of the algorithm has been thoroughly investigated in suicide risk prediction and in the prediction of sub-concepts.

The real time implementation of AFSP within GRiST could have a significant impact on clinical decision making. The algorithm may be used to aid clinical assessment, by providing a risk prediction and comparing the prediction to clinical risk judgements. Since risk predictions are based on consensus, cases in which clinical judgements deviate from consensus may be detected and further investigated. Furthermore, the selected features and their weights could guide clinicians to influential cues and their relative influence for each particular case, and possibly highlight risk areas that need to be managed.

On the other hand, predicting influential sub-concepts; such as current intention of suicide, using AFSP, may be used to alert assessors to the unreliability of a patient's intention statement and/or the assessor's evaluation of intention. In such cases, current intention and suicide risk would need further exploration. AFSP may also be used to predict depression when a clinical diagnosis is not available to aid risk prediction and management.

Finally, an accurate risk prediction algorithm with explanatory capabilities that may be implemented in real time paves the way for self-assessment of mental health

risks. AFSP may also aid risk assessment by para-professionals, who are unable to give their own risk judgements, since it is capable of selecting the most influential features to predict and explain the risk.

Bibliography

[1] S. N. Saleh. *A Novel Dynamic Feature Selection and Prediction Algorithm for Clinical Decisions Involving High-Dimensional and Varied Patient Data*. PhD Thesis, 2016.

[2] Department of Health. Best practice in managing risk: Principles and evidence for best practice in the assessment and management of risk to self and others in mental health services. *https://www.gov.uk/government/publications/assessing-and-managing-risk-in-mental-health-services*, 2009.

[3] M. I Hejazi and X. Cai. Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mmrmr) algorithm. *Advances in water resources*, 32(4):582–593, 2009.

[4] S. E. Hegazy and C. D. Buckingham. A method for automatically eliciting node weights in a hierarchical knowledge based structure for reasoning with uncertainty. *International Journal on Advances in Software*, 2:76–85, 2009.

[5] G. T. Knofczynski and D. Mundfrom. Sample sizes when using multiple linear regression for prediction. *Educational and psychological measurement*, 68(3):431–442, 2008.

[6] W. A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.

[7] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[8] K. Krippendorff. Reliability in content analysis. *Human communication research*, 30(3):411–433, 2004.

[9] National Institute for health and Care Excellence (NICE). *Evidence Standards Framework for Digital Health Technologies*. NICE, 2019.

[10] P. M. Spieth, A. S. Kubasch, A. I. Penzlin, B. M. Illigens, K. Barlinn, and T. Siepmann. Randomized controlled trials–a matter of design. *Neuropsychiatric disease and treatment*, 12:1341, 2016.

[11] World Health Organization. Risks to mental health: An overview of vulnerabilities and risk factors. 2012.

[12] M. J. Prinstein. Introduction to the special section on suicide and nonsuicidal self-injury: A review of unique challenges and important directions for self-injury science. *Journal of consulting and clinical psychology*, 76(1):1, 2008.

[13] World Health Organization. Suicide fact sheet. *http://www.who.int/en/news-room/fact-sheets/detail/suicide*, page accessed June 3rd, 2018.

[14] M. E. Gliatto and A. K. Rai. Evaluation and treatment of patients with suicidal ideation. *American family physician*, 59:1500–1513, 1999.

[15] T.E. Joiner Jr. *Why people die by suicide*. Cambridge, MA: Harvard University Press, 2005.

[16] K. A. Van Orden, T. K. Witte, K. C. Cukrowicz, S. R. Braithwaite, E. A. Selby, and T. E. Joiner Jr. The interpersonal theory of suicide. *Psychological review*, 117(2):575, 2010.

[17] K. Hawton, J. Fagg, S. Platt, and M. Hawkins. Factors associated with suicide after parasuicide in young people. *Bmj*, 309, 1993.

[18] P. Purushothaman, K. C. Premarajan, S. K. Sahu, S. Kattimani, et al. Risk factors and reporting status for attempted suicide: A hospital-based study. *International Journal of Medicine and Public Health*, 5(1):45, 2015.

[19] C. D. Buckingham and A. E. Adams. Classifying clinical decision making: A unifying approach. *Journal of Advanced Nursing*, 32(4):981–989, 2000.

[20] S. Huang, C. Lu, C. Ju, J. Lan, C. Chang, C. L. Chang, W. Kao, C. Lin, and C. Horng. The benefit of clinical psychologists in prevention from the suicide in one hospital in taiwan, republic of china. *Life Science Journal*, 12(3), 2015.

[21] T. Flewett. *Clinical risk management: An introductory text for mental health clinicians*. Elsevier Australia, 2010.

[22] C. D. Buckingham and J. Birtle. Representing the assessment process for psychodynamic psychotherapy within a computerized model of human classification. *British Journal of Medical Psychology*, 70:1–16, 1997.

[23] C. D. Buckingham and D. T. Olsson. Driving information search and retrieval by a cognitive model of classification. In *AMIA 1999 Annual Symposium Conference Proceedings*, Washington DC, 1999.

[24] GRiST. Galatean risk and safety tool. *www.egrist.org*, accessed, May 2018.

[25] C. D. Buckingham, A. Ahmed, and A. Adams. Designing multiple user perspectives and functionality for clinical decision support systems. In *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, pages 211–218. IEEE, 2013.

[26] S. O. Hansson. Risk. *The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), ed. Edward N. Zalta*, 2014.

[27] Y. Y. Haimes. *Risk modeling, assessment, and management*. John Wiley & Sons, 2015.

[28] E. Gilbert, A. Adams, and C. D. Buckingham. Examining the relationship between risk assessment and risk management in mental health. *Journal of Psychiatric and Mental Health Nursing*, 18(10):862–868, 2011.

[29] K. Hawton, C. Casanas, I Comabella, K. Saunders, and C. Haw. Clinical guide: Assessment of suicide risk in people with depression. Technical report, Center for Suicide Research, University of Oxford, 2012.

[30] R. C. Hall, D. E. Platt, and R. C. Hall. Suicide risk assessment: a review of risk factors for suicide in 100 patients who made severe suicide attempts: evaluation of suicide risk in a time of managed care. *Psychosomatics*, 40(1):18–27, 1999.

[31] T. K. Witte, T. E. Joiner, G. K. Brown, A. T. Beck, A. Beckman, P. Duberstein, and Y. Conwell. Factors of suicide ideation and their relation to clinical and other indicators in older adults. *Journal of affective disorders*, 94(1):165–172, 2006.

[32] C. D. Buckingham. Psychological cue use and implications for a clinical decision support system. *Medical Informatics and the Internet in Medicine*, 27(4):237–251, 2002.

[33] M. Lotito and E. Cook. A review of suicide risk assessment instruments and approaches. *Mental Health Clinician*, 5(5):216–223, 2015.

[34] S. Levine, R. J. Ancill, and A. P. Roberts. Assessment of suicide risk by computer-delivered self-rating questionnaire: preliminary findings. *Acta Psychiatrica Scandinavica*, 80(3):216–220, 1989.

[35] W. J. Ferns. Lifenet: a knowledge-based decision support tool for the risk assessment of adolescent suicide. *Expert Systems with Applications*, 9(2):165–176, 1995.

[36] G. S. Brown, G. M. Burlingame, M. J. Lambert, E. Jones, and J. Vaccaro. Pushing the quality envelope: A new outcomes management system. *Psychiatric Services*, 52(7):925–934, 2001.

[37] R. Elzinga and F. Meredith. About face: the applications of a structured approach to mentalhealth information. *Australian Health Review*, 24(1):68–78, 2001.

[38] D. Watts, J. Bindman, M. Slade, F. Holloway, A. Rosen, and G. Thornicroft. Clinical assessment of risk decision support (cards): The development and evaluation of a feasible violence risk assessment for routine psychiatric practice. *Journal of Mental Health*, 13(6):569–581, 2004.

[39] S. A. Webb et al. *Technologies of care*. Jessica Kingsley Publishers Philadelphia, PA, 2003.

[40] R. Shibl, M. Lawley, and J. Debuse. Factors influencing decision support system acceptance. *Decision Support Systems*, 54(2):953–961, 2013.

[41] C. D. Buckingham. Improving mental health risk assessment using web-based decision support. *Health Care Risk Report*, 13(3):17–18, 2007.

[42] C. D. Buckingham and T. Chan. Risk management in mental health. London, 2002. Informa UK.

[43] C. D. Buckingham, A. E. Adams, and C. Mace. Cues and knowledge structures used by mental-health professionals when making risk assessments. *Journal of Mental Health*, 17(3):299–314, 2008.

[44] C. D. Buckingham and A. E. Adams. Employing xml-based technologies to elicit mental-health risk knowledge. In J. Bryant, editor, *Current perspectives in healthcare computing 2006*, pages 284–291. Swindon: BCS HIF, 2006.

[45] C. D. Buckingham, A. Ahmed, and A. E. Adams. Using XML and XSLT for flexible elicitation of mental-health risk knowledge. *Medical Informatics and the Internet in Medicine*, 32(1):65–81, 2007.

[46] C. D. Buckingham, A. Adams, L. Vail, A. Kumar, A. Ahmed, A. Whelan, and E. Karasouli. Integrating service user and practitioner expertise within a web-based system for collaborative mental-health risk and safety management. *Patient Education and Counseling*, 98(10):1189 – 1196, 2015.

[47] C. D. Buckingham and A. Adams. The grist web-based decision support system for mental-health risk assessment and management. In *Proceedings of the First BCS Health in Wales/ehi$_2$ joint Workshop*, pages 37–40, 2011.

[48] H. G. Sanchez. Risk factor model for suicide assessment and intervention. *Professional Psychology: Research and Practice*, 32(4):351, 2001.

[49] M. E. Anderson, M. R. Myhre, D. Suckow, and A. McCabe. Screening and assessment of suicide risk in oncology. *Handbook of Oncology Social Work: Psychosocial Care for People with Cancer*, page 147, 2015.

[50] P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[51] R. Bellman. Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ*, 1961.

[52] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics*. Springer New York, 2017.

[53] C. F. Mela and P. K. Kopalle. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics*, 34(6):667–677, 2002.

[54] C. D. Buckingham and A. E. Adams. Classifying clinical decision making: Interpreting nursing intuition, heuristics, and medical diagnosis. *Journal of Advanced Nursing*, 32(4):990–998, 2000.

[55] T. M. Gale, C. J. Hawley, J. Butler, A. Morton, and A. Singhal. Perception of suicide risk in mental health professionals. *PloS one*, 11(2):e0149791, 2016.

[56] A. Adams, A. Realpe, L. Vail, C. D. Buckingham, L. H. Erby, and D. Roter. How doctors' communication style and race concordance influence african-caribbean patients when disclosing depression. *Patient Education and Counseling*, 98(10):1266–1273, 2015.

[57] S. N. Saleh and C. D. Buckingham. Handling varying amounts of missing data when classifying mental-health risk levels. *Studies in Health Technology and Informatics*, 207:92–101, 2014.

[58] J. R. Cheema. A review of missing data handling methods in education research. *Review of Educational Research*, 84(4):487–508, 2014.

[59] U. Stańczyk and L. C Jain. *Feature selection for data and pattern recognition*. Springer, 2015.

[60] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore. Relief-based feature selection: introduction and review. *Journal of biomedical informatics*, 2018.

[61] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath. Feature subset selection problem using wrapper approach in supervised learning. *International journal of Computer applications*, 1(7):13–17, 2010.

[62] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[63] V. Bolón-Canedo, I. Porto-Díaz, N. Sánchez-Maroño, and A. Alonso-Betanzos. A framework for cost-based feature selection. *Pattern Recognition*, 47(7):2481–2489, 2014.

[64] J. Tang, S. Alelyani, and H. Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014.

[65] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[66] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[67] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Aaai*, volume 2, pages 129–134, 1992.

[68] I. Kononenko, E. Šimec, and M. Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, 7(1):39–55, 1997.

[69] M. Robnik-Šikonja and I. Kononenko. An adaptation of relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*, volume 5, pages 296–304, 1997.

[70] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[71] Y. Sun, S. Todorovic, and S. Goodison. Local-learning-based feature selection for high-dimensional data analysis. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1610–1626, 2010.

[72] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[73] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009.

[74] N. D. Thang, Y. K. Lee, et al. An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information. In *Applications and the Internet (SAINT), 2010 10th IEEE/IPSJ International Symposium on*, pages 395–398. IEEE, 2010.

[75] B. Auffarth, M. López, and J. Cerquides. Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images. In *ICDM*, pages 248–262. Springer, 2010.

[76] Y. Liu, Y. Chen, K. Tan, H. Xie, L. Wang, X. Yan, W. Xie, and Z. Xu. Maximum relevance, minimum redundancy band selection based on neighbor-

hood rough set for hyperspectral data classification. *Measurement Science and Technology*, 27(12):125501, 2016.

[77] J. Clausen. Branch and bound algorithms-principles and examples. *Department of Computer Science, University of Copenhagen*, pages 1–30, 1999.

[78] J. M. Sutter and J. H. Kalivas. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical journal*, 47(1-2):60–66, 1993.

[79] S. J. Russell and Peter N. *Artificial intelligence: a modern approach*. Pearson Education Limited, 2016.

[80] M. Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[81] A. V. Goldberg. Point-to-point shortest path algorithms with preprocessing. In *International Conference on Current Trends in Theory and Practice of Computer Science*. Springer,Berlin, Heidelberg, 2007.

[82] P. Pavel, F. J. Ferri, J. Novovicova, and Kittler J. Floating search methods for feature selection with nonmonotonic criterion functions. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference C: Signal Processing*, volume Vol. 3(2). IEEE, 1994.

[83] A. N. Baraldi and C. K. Enders. An introduction to modern missing data analyses. *Journal of school psychology*, 48(1):5–37, 2010.

[84] S. A. Darabi and B. Teimourpour. A case-based-reasoning system for feature selection and diagnosing asthma. In *Handbook of Research on Data Science for Effective Healthcare Practice and Administration*, pages 444–459. IGI Global, 2017.

[85] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[86] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[87] D. Meyer, F. Leisch, and K. Hornik. The support vector machine under test. *Neurocomputing*, 55(1-2):169–186, 2003.

[88] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. Chang, and L. Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448, 2012.

[89] G. Schwarzer, W. Vach, and M.s Schumacher. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Statistics in medicine*, 19(4):541–561, 2000.

[90] P. R. Harper. A review and comparison of classification algorithms for medical decision making. *Health Policy*, 71(3):315–331, 2005.

[91] J. R. Quinlan. C4.5: Programming for machine learning. *Morgan Kauffmann*, 38:48, 1993.

[92] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[93] R. A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4):597–612, 1922.

[94] C. Kwak and A. Clayton-Matthews. Multinomial logistic regression. *Nursing research*, 51(6):404–410, 2002.

[95] H. F. Yu, F. L. Huang, and C. J. Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, 2011.

[96] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, pages 1470–1480, 1972.

[97] O. Rouhani-Kalleh. Algorithms for fast large scale data mining using logistic regression. In *IEEE Symposium on Computational Intelligence and Data Mining*, pages 155–162. IEEE, 2007.

[98] G. King and L. Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[99] S. Dias, A. E. Ades, N. J. Welton, J. P. Jansen, and A. J. Sutton. *Network Meta-Analysis for Decision Making*. Wiley Online Library, 2018.

[100] A. J. Dobson and A. Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2008.

[101] NCSS. Statistical software, ridge regression, chapter 335. *http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf*, accessed April 2019.

[102] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[103] G. A. Fink. *Markov Models for Pattern Recognition*. Springer, 2014.

[104] S. Günter and H. Bunke. Optimizing the number of states, training iterations and gaussians in an hmm-based handwritten word recognizer. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 472–476. IEEE, 2003.

[105] I. Miklós and I. M. Meyer. A linear memory algorithm for baum-welch training. *BMC bioinformatics*, 6(1):231, 2005.

[106] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[107] K. Chan, T. W. Lee, and T. J. Sejnowski. Handling missing data with variational bayesian learning of ica. In *Advances in Neural Information Processing Systems*, pages 905–912, 2003.

[108] O. Obembe and C. D. Buckingham. Developing a probabilistic graphical structure from a model of mental-health clinical risk expertise. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6279 LNAI(PART 4):88–97, 2010.

[109] Olufunmilayo O. *Development of a Probabilistic Graphical Structure from a Model of Mental Health Clinical Expertise*. PhD Thesis, 2011.

[110] J. S. Long and L. H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[111] K. Iyer. *Computational Complexity of Data Mining Algorithms Used in Fraud Detection*. PhD Thesis, 2015.

[112] R. J. Carroll. *Transformation and weighting in regression*. Routledge, 2017.

[113] T. D. Wickens. *Elementary signal detection theory*. Oxford University Press, USA, 2002.

[114] A. Agresti and M. Kateri. *Categorical data analysis*. Springer, 2011.

[115] K. Huang, H. Yang, I. King, M. R. Lyu, and L. Chan. The minimum error minimax probability machine. *The Journal of Machine Learning Research*, 5:1253–1286, 2004.

[116] N. A. Zaher and C. D. Buckingham. Moderating the influence of current intention to improve suicide risk prediction. In *AMIA annual symposium proceedings*, pages 1274–1285. American Medical Informatics Association, 2016.

[117] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, et al. *Numerical recipes*, volume 3. cambridge University Press, cambridge, 1989.

[118] J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506, 1953.

[119] D. Owens, J. Horrocks, and A. House. Fatal and non-fatal repetition of self-harm: systematic review. *The British Journal of Psychiatry*, 181(3):193–199, 2002.

[120] S. T. Adams and Stephen H. L. Clinical prediction rules. *Bmj*, 344, 2012, d8312.

[121] Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[122] J. B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[123] S. Weisborg. *Applied Linear Regression*, volume 528. John Wiley & Sons, 2005.

[124] T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294, 1979.

[125] J. C. Franklin, J. D. Ribeiro, K. R. Fox, K. H. Bentley, E. M. Kleiman, X. Huang, K. M. Musacchio, A. C. Jaroszewski, B. P. Chang, and M. K. Nock. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187, 2017.

[126] L. Harriss, K. Hawton, and D. Zahl. Value of measuring suicidal intent in the assessment of people attending hospital following self-poisoning or self-injury. *The British Journal of Psychiatry*, 186(1):60–66, 2005.

[127] S. Schächtele, T. Gerstenberg, and D. Lagnado. Beyond outcomes: The influence of intentions and deception. In *Proceedings of the 33nd annual conference of the cognitive science society. Cognitive Science Society*, pages 1860–1865, 2011.

[128] M. Bratman. Intention, plans, and practical reason. 1987.

[129] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial intelligence*, 42(2):213–261, 1990.

[130] D. L. Roter, L. H. Erby, A. Adams, C. D. Buckingham, L. Vail, A. Realpe, S. Larson, and J. A. Hall. Talking about depression: An analogue study of physician

gender and communication style on patient disclosures. *Patient Education and Counseling*, 96(3):339–345, 2014.

[131] T. A. Mieczkowski, J. A. Sweeney, G. L. Haas, B. W. Junker, R. P. Brown, and J. J. Mann. Factor composition of the suicide intent scale. *Suicide and Life-Threatening Behavior*, 23(1):37–45, 1993.

[132] M. Fava and K. S. Kendler. Major depressive disorder. *Neuron*, 28(2):335–341, 2000.

[133] Y. J. Pan, K. D. Juang, S. R. Lu, S. P. Chen, Y. F. Wang, J. L. Fuh, and S. J. Wang. Longitudinal risk factors for suicidal thoughts in depressed and non-depressed young adolescents. *Australian & New Zealand Journal of Psychiatry*, 51(9):930–937, 2017.

[134] D. R. Jahn, K. C. Cukrowicz, K. Linton, and F. Prabhu. The mediating effect of perceived burdensomeness on the relation between depressive symptoms and suicide ideation in a community sample of older adults. *Aging & Mental Health*, 15(2):214–220, 2011.

[135] A. Ahmed. *Knowledge engineering for mental-health risk assessment and decision support*. PhD Thesis, 2011.

[136] I. D'Haeseleer, B. Vanrumste, D. Schreurs, C. D. Buckingham, A. Mondelaers, and V. V. Abeele. Attitudes of older adults towards self-assessment of mental health, safety and wellbeing. *Engineering4Society 2016*, 1:73, 2016.

[137] T. Greenhalgh, N. Fahy, and S. Shaw. The bright elusive butterfly of value in health technology development: Comment on" providing value to new health technology: The early contribution of entrepreneurs, investors, and regulatory agencies". *International journal of health policy and management*, 7(1):81, 2018.

[138] K. P. Suresh. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of human reproductive sciences*, 4(1):8, 2011.

[139] S. Johansen and B. Nielsen. Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, 43(2):321–348, 2016.